Technical Report 1306

Tier One Performance Screen Initial Operational Test and Evaluation: 2011 Interim Report

Deirdre J. Knapp, Editor
Human Resources Research Organization

Tonia S. Heffner, Editor
U.S. Army Research Institute

April 2012

United States Army Research Institute for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS, Ph.D.
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Nehama Babin, U.S. Army Research Institute
Tatana Olson, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 6000 6th Street, Bldg 1464 / Mail Stop 5610, Fort Belvoir, VA 22060-5610.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
April 2012	Interim	August 2009 to May 2011

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Tier One Performance Screen Initial Operational Test and Evaluation: 2011 Interim Report	W91WAS-09-D-0013
	5b. PROGRAM ELEMENT NUMBER: 622785
6. AUTHOR(S): Deirdre J. Knapp (Ed.) (Human Resources Research Organization), and Tonia S. Heffner (Ed.) (U.S. Army Research Institute)	5c. PROJECT NUMBER: A790
	5d. TASK NUMBER: 329
	5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Human Resources Research Organization 66 Canal Center Plaza, Suite 700 Alexandria, Virginia 22314	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences ATTN: DAPE-ARI 6000 6th Street (Bldg. 1464 / Mail Stop 5610) Fort Belvoir, VA 22060-5610	10. MONITOR ACRONYM: ARI
	11. MONITOR REPORT NUMBER: Technical Report 1306

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contracting Officer's Representative and Subject Matter Expert POC: Kate LaPort

14. ABSTRACT *(Maximum 200 words)*:

Along with educational, medical, and moral screens, the U.S. Army uses a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new Soldiers. Although the AFQT is useful for selecting new Soldiers, other personal attributes are important to Soldier performance and retention. Based on the U.S. Army Research Institute's (ARI) investigations, the Army selected one promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), for an initial operational test and evaluation (IOT&E), beginning administration to applicants in 2009. Criterion data are being compiled at 6-month intervals from administrative records, from Initial Military Training (IMT), and from schools for eight military occupational specialties (MOS) and will be followed by multiple waves of data collection from Soldiers in units. This is the third of six planned evaluations of the TAPAS. Similar to prior research, results thus far suggest that several TAPAS scales significantly predict a number of criteria of interest, indicating that the measure holds promise for both selection and classification purposes.

15. SUBJECT TERMS

Personnel, Manpower, Selection and classification

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	95	Ellen Kinzer Technical Publications Specialist (703) 545-4225

Standard Form 298

Technical Report 1306

Tier One Performance Screen Initial Operational Test and Evaluation: 2011 Interim Report

Deirdre J. Knapp, Editor
Human Resources Research Organization

Tonia S. Heffner, Editor
U.S. Army Research Institute

Personnel Assessment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
6000 6th Street, Bldg 1464
Fort Belvoir, VA 22060

April 2012

Army Project Number
622785A790

Personnel, Performance and Training Technology

ACKNOWLEDGEMENTS

There are individuals not listed as authors who made significant contributions to the research described in this report. First and foremost are the Army cadre who support criterion data collection efforts at the schoolhouses. These noncommissioned officers (NCOs) ensure that trainees are scheduled to take the research measures and provide ratings of their Soldiers' performance in training. Thanks also go to Ms. Kate LaPort, Mr. Irwin Jose, and Ms. Sharon Meyers (ARI) and Mr. Doug Brown, Ms. Ashley Armstrong, and Ms. Mary Adeniyi (HumRRO) and Mr. Jason Vetter (Drasgow Consulting Group) for their contributions to this research effort.

We also want to extend our appreciation to the Army Test Program Advisory Team (ATPAT), a group of senior NCOs who periodically meet with ARI researchers to help guide this work in a manner that ensures its relevance to the Army and help enable the Army support required to implement the research. Members of the ATPAT are listed below:

MAJ PAUL WALTON
CSM BRIAN A. HAMM
CSM JAMES SHULTZ
SGM KENAN HARRINGTON
SGM THOMAS KLINGEL
SGM(R) CLIFFORD MCMILLAN
SGM GREGORY A. RICHARDSON
SFC WILLIAM HAYES
SFC KENNETH WILLIAMS
MR. ROBERT STEEN

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: 2011 INTERIM REPORT

EXECUTIVE SUMMARY

Research Requirement:

In addition to educational, physical, and moral screens, the U.S. Army relies on a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new Soldiers into the Army. Although the AFQT has proven to be, and will continue to serve as, a useful metric for selecting new Soldiers, other personal attributes, in particular non-cognitive attributes (e.g., temperament, interests, and values), are important to entry-level Soldier performance and retention (e.g., Campbell & Knapp, 2001; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009, 2010; Knapp & Tremble, 2007). Based on ARI's research, the Army selected one particularly promising measure, the Tailored Adaptive Personality Assessment System (TAPAS), as the basis for an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen* (TOPS). The TAPAS capitalizes on the latest in testing technology to assess motivation through the measurement of personality characteristics.

In May 2009, the Military Entrance Processing Command (MEPCOM) began administering the TAPAS on the computer adaptive platform for the ASVAB (CAT-ASVAB) at Military Entrance Processing Stations (MEPS). The Work Preferences Assessment (WPA), which asks respondents their preference for various work activities and environments, will also be introduced for applicant testing in CY2012. Both measures will be administered as part of the IOT&E through FY2013. The Information/Communication Technology Literacy (ICTL) test developed by the Air Force is being administered to a subset of applicants as part of the IOT&E as of FY2011. Criterion data are being compiled from administrative records at 6-month intervals. As part of the IOT&E, initial military training (IMT) criterion data are currently being collected at schools for eight military occupational specialties (MOS) and the first of multiple waves of data collection from Soldiers in their units has been initiated.

Procedure:

To evaluate the TAPAS, ICTL, and WPA, the Army is collecting training criterion data on Soldiers in selected MOS as they complete their IMT. The criterion measures include job knowledge tests (JKTs); an attitudinal assessment, the Army Life Questionnaire (ALQ); and performance rating scales (PRS) completed by the Soldiers' cadre members. Course grades, completion rates, and attrition status are obtained from administrative records for all Soldiers, regardless of MOS.

The May 2011 data file, which was the basis for analyses documented in this report, includes a total of 151,625 applicants who took the TAPAS, 141,483 of whom were in the TOPS "Applicant Sample." The Applicant Sample used for analysis purposes excluded Education Tier 3, AFQT Category V, and prior service applicants. The validation sample sizes are considerably

smaller, with the Schoolhouse Validation Sample comprising 4,976 Soldiers and the Validation Sample (which includes Soldiers for whom we only have administrative criterion data) comprising 46,188 Soldiers.

Our approach to analyzing the TAPAS' incremental predictive validity was consistent with previous evaluations of this measure and similar experimental non-cognitive predictors (Ingerick et al., 2009; Knapp & Heffner, 2009, 2010, 2011). In brief, this approach involved testing a series of hierarchical regression models, regressing each criterion measure onto Soldiers' AFQT scores in the first step, followed by their TAPAS scale scores in the second step. When the TAPAS scale scores were added to the baseline regression models, the resulting increment in the multiple correlation (ΔR) served as our index of incremental validity. In the present research, new TAPAS composites were also formed, relying on metrics of relative importance such as regression weights and relative weights. When re-scaled to a proportion metric ranging from 0.0-100.0%, relative weights can be interpreted as the percentage of criterion variance accounted for (R^2) by each TAPAS scale.

Similar to previous research (Ingerick et al. 2009; Knapp, Owens, & Allen, 2011), we evaluated the experimental predictor measures' classification potential using (a) Horst's (1954, 1955) index of differential validity (H_d) and (b) Brogden's expected criterion scores of optimally assigned individuals (De Corte, 2000). We also examined incremental classification beyond ASVAB.

Findings:

Consistent with previous analyses in the TOPS stream of research (Caramagno, Allen, & Ingerick, 2011; Trippe, Caramagno, Allen, & Ingerick, 2011), results suggest that the TAPAS holds promise for predicting key criteria of interest. Incremental validity beyond the ASVAB is reasonably strong, especially for will-do criterion measures (i.e., those measuring non-technical aspects of Soldier performance, such as effort, peer leadership, and personal discipline). This is despite the low reliability of the supervisor ratings.

Multiple approaches were used to develop alternative TOPS composites. Due to their nature, results of these analyses are reported separately in a set of two limited distribution appendices.[1] Analyses conducted to evaluate these alternative composites show that they outperform the original composites developed in the experimental phase of the project (Knapp & Heffner, 2010) in terms of predictive utility.

The classification results presented in the present report provide further evidence that the TAPAS can provide incremental improvements beyond the ASVAB subtests for optimal assignment of Soldiers to MOS. These incremental gains were observed in the dichotomous outcome variables (attrition and IMT restart). In these cases, the reduction in predicted overall attrition or IMT restart is modest, but some MOS level results suggest a significant improvement. This is to some extent a result of the classification scenario modeled here, in which every Soldier

[1] Contact editors for availability of appendices describing alternative TOPS composite development and validation.

must be assigned to an MOS according to the allocation percentages. Results from the JKT and ALQ variables currently do not suggest that the TAPAS provides incremental improvements in
classification beyond the ASVAB for these criteria, but those analyses were based on far smaller sample sizes than the attrition and IMT restart analyses. In other words, the analyses using "for-research-only" criteria suffer from a number of limitations related to the availability of criterion data within and across the MOS. We expect these analyses to become more informative as these criterion data continue to accumulate and we obtain JKT and ALQ data for additional MOS.

Taken together, evaluation results thus far suggest that, while the magnitude of the validity and classification coefficients are not as large as those found in the experimental *Expanded Enlistment Eligibility Metrics* (EEEM) research (Knapp & Heffner, 2010), the TAPAS holds promise for both selection and classification purposes. Many of the scale-level coefficients are consistent with a theoretical understanding of the TAPAS scales, suggesting that the scales are measuring the characteristics that they are intended to measure. However, given the restricted nature of the matched criterion sample (in terms of sample characteristics) and the low reliability of the ratings data, these results should be considered preliminary.

Utilization and Dissemination of Findings:

The research findings will be used by the, U.S. Army Recruiting Command, Army G-1, and Training and Doctrine Command to evaluate the effectiveness of tools used for Army applicant selection and assignment. With each successive set of findings, the TOPS can be revised and refined to meet Army needs and requirements.

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: 2011 INTERIM REPORT

CONTENTS

Page

CHAPTER 1: INTRODUCTION....1
Deirdre J. Knapp (HumRRO), Tonia S. Heffner, and Leonard A. White (ARI)
Background 1
The Tier One Performance Screen (TOPS).... 2
Evaluating TOPS.... 3
Overview of Report.... 4

CHAPTER 2: DATA FILE DEVELOPMENT....5
D. Matthew Trippe, Laura Ford, Bethany Bynum, and Karen Moriarty (HumRRO)
Overview of Process.... 5
Description of Data File and Sample Construction.... 6
Summary 10

CHAPTER 3: DESCRIPTION OF THE TOPS IOT&E PREDICTOR MEASURES....11
Stephen Stark, O. Sasha Chernyshenko, Fritz Drasgow (Drasgow Consulting Group), and Matthew T. Allen (HumRRO)
Tailored Adaptive Personality Assessment System (TAPAS).... 11
TAPAS Background 11
Three Current Versions of TAPAS....12
TAPAS Scoring 14
TAPAS Initial Validation Effort....16
Initial TAPAS Composites 17
Armed Services Vocational Aptitude Battery (ASVAB) Content, Structure, and Scoring . 18
Summary 19

CHAPTER 4: DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF CRITERION MEASURES 20
Karen O. Moriarty and Tina Chang (HumRRO)
Training Criterion Measure Descriptions.... 21
Job Knowledge Tests (JKTs)....21
Performance Rating Scales (PRS) 21
Army Life Questionnaire (ALQ) 23
Administrative Criteria 23
Training Criterion Measure Scores and Associated Psychometric Properties 23
Job Knowledge Tests (JKTs)....24
Performance Rating Scales (PRS) 25
Army Life Questionnaire (ALQ) 27
Administrative Criterion Data....28
Summary 29

CONTENTS (contined)

Page

CHAPTER 5: VALIDITY RESULTS AND COMPOSITE FORMATION 30
Matthew C. Reeder, Matthew T. Allen, and Michael J. Ingerick (HumRRO)
Background and Approach 30
Evaluating the Predictive Potential of the TAPAS 30
Predictive Potential of the TAPAS in the Full Validation Sample 30
Predictive Potential of TAPAS Scales in the Restricted (AFQT Category IIIB and IV) Sample 33
Constructing and Evaluating Revised TOPS Composites 33
Summary 35

CHAPTER 6: EVALUATION OF TAPAS POTENTIAL FOR CLASSIFICATION PURPOSES 36
Matthew Trippe, Ted Diaz, and Michael Ingerick (HumRRO)
Introduction 36
Approach to Estimating the Classification Potential 36
Results 38
Summary 45

CHAPTER 7: SUMMARY AND A LOOK AHEAD 46
Deirdre J. Knapp (HumRRO), Tonia S. Heffner, and Leonard A. White (ARI)
Summary of the TOPS IOT&E Method 46
Summary of Evaluation Results to Date 46
TAPAS Construct Validity 46
Validity for Soldier Selection 47
Potential for Soldier Classification 47
Results Summary 47
Looking Ahead 48
Predictor Measures 48
Criterion Measures 49
In-Unit Data Collections 49
Analyses 49

REFERENCES 51

APPENDIX A: PREDICTOR MEASURE PSYCHOMETRIC PROPERTIES A-1

APPENDIX B: CRITERION MEASURE PSYCHOMETRIC PROPERTIES IN FULL SCHOOLHOUSE SAMPLE B-1

APPENDIX C: VALIDITY SUPPORTING ANALYSES C-1

Page

LIST OF TABLES

TABLE 2.1. FULL TAPAS DATA FILE SAMPLE CHARACTERISTICS 7
TABLE 2.2. DISTRIBUTION OF MOS IN THE FULL SCHOOLHOUSE DATA FILE 8
TABLE 2.3. BACKGROUND AND DEMOGRAPHIC CHARACTERISTICS OF THE TOPS SAMPLES 9

TABLE 3.1. TAPAS DIMENSIONS ASSESSED 13

TABLE 4.1. SUMMARY OF TRAINING CRITERION MEASURES 20
TABLE 4.2. EXAMPLE TRAINING PERFORMANCE RATING SCALES 21
TABLE 4.3. ALQ LIKERT-TYPE SCALES 24
TABLE 4.4. DESCRIPTIVE STATISTICS AND RELIABILITY ESTIMATES FOR TRAINING JOB KNOWLEDGE TESTS (JKTS) IN THE VALIDATION SAMPLE 25
TABLE 4.5. DESCRIPTIVE STATISTICS AND RELIABILITY ESTIMATES FOR TRAINING PERFORMANCE RATING SCALES (PRS) IN THE VALIDATION SAMPLE 26
TABLE 4.6. DESCRIPTIVE STATISTICS AND RELIABILITY ESTIMATES FOR THE ARMY LIFE QUESTIONNAIRE (ALQ) IN THE TOPS VALIDATION SAMPLE 27
TABLE 4.7. DESCRIPTIVE STATISTICS FOR ADMINISTRATIVE CRITERIA IN THE TOPS VALIDATION SAMPLE 28

TABLE 5.1. INCREMENTAL VALIDITY ESTIMATES FOR THE TAPAS SCALES OVER THE AFQT FOR PREDICTING PERFORMANCE- AND RETENTION-RELATED CRITERIA 32
TABLE 5.2. VALIDITY ESTIMATES FOR THE TAPAS SCALES FOR PREDICTING PERFORMANCE- AND RETENTION-RELATED CRITERIA FOR EDUCATION TIER 1, AFQT CATEGORY IIIB AND IV SOLDIERS 34

TABLE 6.1. SUMMARY OF OVERALL CLASSIFICATION POTENTIAL INDICES OF THE TAPAS RELATIVE TO THE ASVAB BY CRITERION MEASURE 39
TABLE 6.2. CLASSIFICATION POTENTIAL OF THE TAPAS RELATIVE TO THE ASVAB FOR MINIMIZING 3-MONTH ATTRITION 39
TABLE 6.3. CLASSIFICATION POTENTIAL OF THE TAPAS RELATIVE TO THE ASVAB FOR MINIMIZING 6-MONTH ATTRITION 42
TABLE 6.4. CLASSIFICATION POTENTIAL OF THE TAPAS RELATIVE TO THE ASVAB FOR MINIMIZING IMT RESTART 43

CONTENTS (CONTINUED)

Page

TABLE 6.5. CLASSIFICATION POTENTIAL OF THE TAPAS RELATIVE TO THE ASVAB FOR MAXIMIZING MOS SPECIFIC JKT SCORES 43
TABLE 6.6. CLASSIFICATION POTENTIAL OF THE TAPAS RELATIVE TO THE ASVAB FOR MAXIMIZING PERCEPTIONS OF ARMY LIFE ADJUSTMENT AND MOS FIT 44

TABLE 7.1. TAPAS DIMENSIONS ASSESSED 48

TABLE A.1. RAW MEANS AND STANDARD DEVIATIONS FOR THE TOPS IOT&E TAPAS SCALES BY VERSION A-2
TABLE A.2. TAPAS FACET SCALE AND AFQT INTERCORRELATIONS A-3
TABLE A.3. DESCRIPTIVE STATISTICS FOR THE ASVAB A-4
TABLE A.4. ASVAB SUBSET AND AFQT INTERCORRELATIONS A-5
TABLE A.5. RACE, ETHNIC, AND GENDER SUBGROUP MEANS AND STANDARD DEVIATIONS A-6
TABLE A.6. RACE, ETHNIC, AND GENDER SUBGROUP DIFFERENCES IN MEANS A-7

TABLE B.1. DESCRIPTIVE STATISTICS FOR TRAINING CRITERIA BASED ON THE FULL SCHOOLHOUSE SAMPLE B-2
TABLE B.2. DESCRIPTIVE STATISTICS FOR SCHOOLHOUSE CRITERIA BY MOS FROM THE FULL SCHOOLHOUSE SAMPLE B-3
TABLE B.3. ARMY LIFE QUESTIONNAIRE (ALQ) INTERCORRELATIONS FOR FULL SCHOOLHOUSE SAMPLE B-4
TABLE B.4. PERFORMANCE RATING SCALES (PRS) INTERCORRELATIONS FOR FULL SCHOOLHOUSE SAMPLE B-4
TABLE B.5. CORRELATIONS BETWEEN THE ARMY LIFE QUESTIONNAIRE (ALQ) AND JOB KNOWLEDGE TESTS (JKT) IN FULL SCHOOLHOUSE SAMPLE B-5
TABLE B.6. CORRELATIONS BETWEEN THE ARMY LIFE QUESTIONNAIRE (ALQ) AND PERFORMANCE RATING SCALES (PRS) IN FULL SCHOOLHOUSE SAMPLE B-5
TABLE B.7. CORRELATIONS BETWEEN JOB KNOWLEDGE TESTS (JKT) AND PERFORMANCE RATING SCALES (PRS) IN FULL SCHOOLHOUSE SAMPLE B-6
TABLE B.8. DESCRIPTIVE STATISTICS FOR ADMINISTRATIVE CRITERIA BASED ON THE TOPS VALIDATION SAMPLE BY MOS B-7

Page

TABLE B.9. CORRELATIONS BETWEEN THE ARMY LIFE QUESTIONNAIRE (ALQ) CRITERIA AND ADMINISTRATIVE CRITERIA B-8

TABLE B.10. CORRELATIONS BETWEEN THE PERFORMANCE RATING SCALES (PRS) AND ADMINISTRATIVE CRITERIA B-9

TABLE B.11. CORRELATIONS BETWEEN RESEARCH ONLY AND ADMINISTRATIVE CRITERIA B-9

TABLE C.1. BIVARIATE CORRELATIONS BETWEEN THE TAPAS SCALES AND CAN-DO PERFORMANCE-RELATED CRITERIA FOR TIER 1 SOLDIERS C-2

TABLE C.2. BIVARIATE CORRELATIONS BETWEEN THE TAPAS SCALES AND WILL-DO PERFORMANCE-RELATED CRITERIA FOR TIER 1 SOLDIERS C-3

TABLE C.3. BIVARIATE CORRELATIONS BETWEEN THE TAPAS SCALES AND RETENTION-RELATED CRITERIA FOR TIER 1 SOLDIERS C-4

TABLE C.4. BIVARIATE CORRELATIONS BETWEEN THE TAPAS SCALES AND CAN-DO PERFORMANCE-RELATED CRITERIA FOR TIER 2 SOLDIERS C-5

TABLE C.5. BIVARIATE CORRELATIONS BETWEEN THE TAPAS SCALES AND WILL-DO PERFORMANCE-RELATED CRITERIA FOR TIER 2 SOLDIERS C-6

TABLE C.6. BIVARIATE CORRELATIONS BETWEEN THE TAPAS SCALES AND RETENTION-RELATED CRITERIA FOR TIER 2 SOLDIERS C-7

LIST OF FIGURES

FIGURE 1.1. TOPS INITIAL OPERATIONAL TEST & EVALUATION (IOT&E). 3

FIGURE 2.1. SUMMARY OF TOPS SCHOOLHOUSE (IMT) DATA SOURCES. 5

FIGURE 2.2. OVERVIEW OF TOPS DATA FILE MERGING AND NESTED SAMPLE GENERATION PROCESS 6

FIGURE 4.1. SAMPLE 7-POINT BEHAVIORALLY-ANCHORED RATING SCALE 22

FIGURE 4.2. RELATIVE OVERALL PERFORMANCE RATING SCALE 22

FIGURE 6.1. BOXPLOT PRESENTATION OF THE CLASSIFICATION POTENTIAL OF THE TAPAS RELATIVE TO THE ASVAB FOR MINIMIZING 3-MONTH ATTRITION 41

TIER ONE PERFORMANCE SCREEN INITIAL OPERATIONAL TEST AND EVALUATION: 2011 INTERIM REPORT

CHAPTER 1: INTRODUCTION

Deirdre J. Knapp (HumRRO), Tonia S. Heffner, and Leonard A. White (ARI)

Background

The Personnel Assessment Research Unit (PARU) of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is responsible for conducting personnel research for the Army. The focus of PARU's research is maximizing the potential of the individual Soldier through effective selection, classification, and retention strategies.

In addition to educational, physical, and moral screens, the U.S. Army relies on a composite score from the Armed Services Vocational Aptitude Battery (ASVAB), the Armed Forces Qualification Test (AFQT), to select new Soldiers into the Army. Although the AFQT has proven to be, and will continue to serve as, a useful metric for selecting new Soldiers, other personal attributes, in particular non-cognitive attributes (e.g., temperament, interests, and values), are important to entry-level Soldier performance and retention (e.g., Knapp & Tremble, 2007).

In December 2006, the Department of Defense (DoD) ASVAB review panel—a panel of experts in the measurement of human characteristics and performance— released their recommendations (Drasgow, Embretson, Kyllonen, & Schmitt, 2006). Several of these recommendations focused on supplementing the ASVAB with additional measures for use in selection and classification decisions. The ASVAB review panel further recommended that the use of these measures be validated against performance criteria.

Just prior to release of the ASVAB review panel's findings, ARI had initiated a longitudinal research effort, *Validating Future Force Performance Measures (Army Class)*, to examine the prediction potential of several non-cognitive measures (e.g., temperament and person-environment fit) for Army outcomes (e.g., performance, attitudes, attrition). The Army Class research project is a 6-year effort that is being conducted with contract support from the Human Resources Research Organization ([HumRRO]; Ingerick, Diaz, & Putka, 2009; Knapp & Heffner, 2009). Experimental predictors were administered to new Soldiers in 2007 and early 2008. Since then, Army Class researchers have obtained attrition data from Army records and collected training criterion data on a subset of the Soldier sample. Job performance criterion data were collected from Soldiers in the Army Class longitudinal validation sample in 2009 (Knapp, Owens, & Allen, 2011) and a second round of job performance data collection was completed in April 2011.

After the Army Class research was underway, ARI initiated the *Expanded Enlistment Eligibility Metrics (EEEM)* project (Knapp & Heffner, 2010). The EEEM goals were similar to Army Class, but the focus was specifically on Soldier selection (as opposed to selection and Military Occupational Specialty [MOS] classification) and the time horizon was much shorter. Specifically, EEEM required selection of one or more promising new predictor measures for immediate implementation. The EEEM project capitalized on the existing Army Class data collection procedure and, thus, the EEEM sample was a subset of the Army Class sample.

As a result of the EEEM findings, Army policy-makers approved an initial operational test and evaluation (IOT&E) of the *Tier One Performance Screen (TOPS)*. This report is the third in a series presenting analyses from the IOT&E of TOPS.

The Tier One Performance Screen (TOPS)

Six experimental pre-enlistment measures were included in the EEEM research (Allen, Cheng, Putka, Hunter, & White, 2010). These included several temperament measures, a situational judgment test, and two person-environment fit measures based on values and interests. The "best bet" measures recommended to the Army for implementation were identified based on the following considerations:

- Incremental validity over AFQT for predicting important performance and retention-related outcomes
- Minimal subgroup differences
- Low susceptibility to response distortion (e.g., faking good)
- Minimal administration time requirements

The Tailored Adaptive Personality Assessment System ([TAPAS]; Stark, Chernyshenko, & Drasgow, 2010b) surfaced as the top choice, with the Work Preferences Assessment ([WPA]; Putka & Van Iddekinge, 2007) identified as another good option that was substantively different from the TAPAS. Specifically, TAPAS is a measure of personality characteristics (e.g., achievement, sociability) that capitalizes on the latest advances in psychometric theory and provides a good indicator of personal motivation. The WPA asks applicants to indicate their preference for various kinds of work activities and environments (e.g., "A job that requires me to teach others," "A job that requires me to work outdoors"). Although not included in the EEEM research, the Information/Communications Technology Literacy (ICTL) test emerged as a potential test of applicants' familiarity with computers and information technology, which may predict performance in high-technology occupations.

In May 2009, the Military Entrance Processing Command (MEPCOM) began administering TAPAS on the computer adaptive platform for the ASVAB (CAT-ASVAB). Initially, TAPAS was to be administered only to Education Tier 1 which are primarily high school diploma graduates, non-prior service applicants.[2] This limitation was removed several months after the start so the Army could evaluate TAPAS across all types of applicants. The TAPAS administration by MEPCOM is scheduled to continue through the fall of 2012.

TOPS uses non-cognitive measures to identify Education Tier 1 applicants who would likely perform differently (higher or lower) than would be predicted by their ASVAB scores. As part of the TOPS IOT&E, TAPAS scores are being used to screen out a small number of AFQT Category IIIB/ IV applicants.[3] Although the WPA is part of the TOPS IOT&E, WPA scores will not be considered

[2] Applicant educational credentials are classified as Tier 1 (high school diploma), Tier 2 (non-diploma graduate), and Tier 3 (not a high school graduate).

[3] Examinees are classified into categories based on their AFQT percentile scores (Category I = 93-99, Category II = 65-92, Category IIIA = 50-54, Category IIIB = 31-49, Category IV = 10-30, Category V = 1-9).

for enlistment eligibility. The WPA is being prepared for MEPS administration with an expected start date of May 2012.

Although the initial conceptualization for the IOT&E was to use TAPAS as a tool for "screening in" Education Tier 1 applicants with lower AFQT scores, changing economic conditions spurred a reconceptualization to a system that screens out low motivated applicants with low AFQT scores. It is likely that the selection model in a fully operational system would adjust to fit with the changing applicant market. For example, at the present time, few applicants are being screened out based on TAPAS scores, not just because the passing scores are set quite low, but also because there are very few Category IV applicants being considered for enlistment due to the overwhelming availability of applicants in higher AFQT categories. Because many factors may impact how TAPAS would be used in the applicant screening process, TAPAS is currently administered to all Education Tier 1 and Tier 2 non-prior service applicants who take the ASVAB at the MEPS.

Evaluating TOPS

Figure 1.1 illustrates the TOPS IOT&E research plan. To evaluate the non-cognitive measures (TAPAS and WPA), the Army is collecting training criterion data on Soldiers in eight target military occupational specialties (MOS) as they complete initial military training (IMT).[4] The criterion measures include job knowledge tests (JKTs); an attitudinal assessment, the Army Life Questionnaire (ALQ); and performance rating scales (PRS) completed by the Soldiers' cadre. These measures are computer-administered at the schools for each of the eight target MOS. The process is overseen by Army personnel with guidance and support from both ARI and HumRRO. Course grades and completion rates are obtained from administrative records for all Soldiers who take the TAPAS, regardless of MOS.

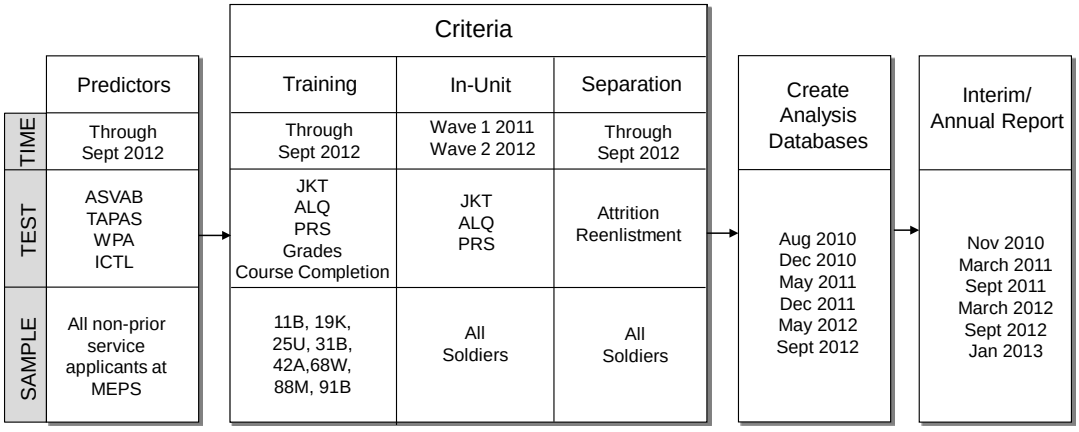

Figure 1.1. TOPS Initial Operational Test & Evaluation (IOT&E).

Two waves of in-unit job performance data collection are also planned, both of which will attempt to include Soldiers from across all MOS who completed the TAPAS (and WPA)

[4] The target MOS are Infantryman (11B), Armor Crewman (19K), Signal Support Specialist (25U), Military Police (31B), Human Resources Specialist (42A), Health Care Specialist (68W), Motor Transport Operator (88M), and Light Wheel Vehicle Mechanic (91B). These MOS were selected to include large, highly critical MOS as well as to represent the diversity of work requirements across MOS.

during the application process. These measures will again include JKTs, the ALQ, and cadre ratings. Finally, the separation status of all Soldiers who took the TAPAS is being tracked throughout the course of the research.

This report describes the third iteration to develop a criterion-related validation data file and conduct evaluation analyses using data collected in the TOPS IOT&E initiative. Prior evaluations are described in Knapp, Heffner, and White (2011) and Knapp and Heffner (2011). Additional analysis datasets and validation analyses will be prepared and conducted at 6-month intervals throughout the multi-year IOT&E period.

Overview of Report

Chapter 2 explains how the evaluation analysis data files are constructed, then describes characteristics of the samples resulting from construction of the latest analysis data file in May 2011. Chapter 3 describes the TAPAS and ASVAB, including content, scoring, and psychometric characteristics. Chapter 4 describes the criterion measures included in this analysis, including their psychometric characteristics. Criterion-related validation analyses are presented in Chapter 5. Chapter 6 presents analyses examining the classification potential of the TAPAS. The report concludes with Chapter 7, which summarizes our continuing efforts to evaluate TOPS and looks toward plans for future iterations of these evaluations.

CHAPTER 2: DATA FILE DEVELOPMENT

D. Matthew Trippe, Laura Ford, Bethany Bynum, and Karen Moriarty (HumRRO)

Overview of Process

The TOPS data file is assembled from a number of sources. In general, the data file comprises predictor and criterion data obtained from administrative and IMT (or "schoolhouse") sources.[5] IMT records comprise assessment data collected from Soldiers and their cadre (i.e., supervisors) at the locations identified in Figure 2.1. The IMT assessments were developed specifically for this IOT&E.

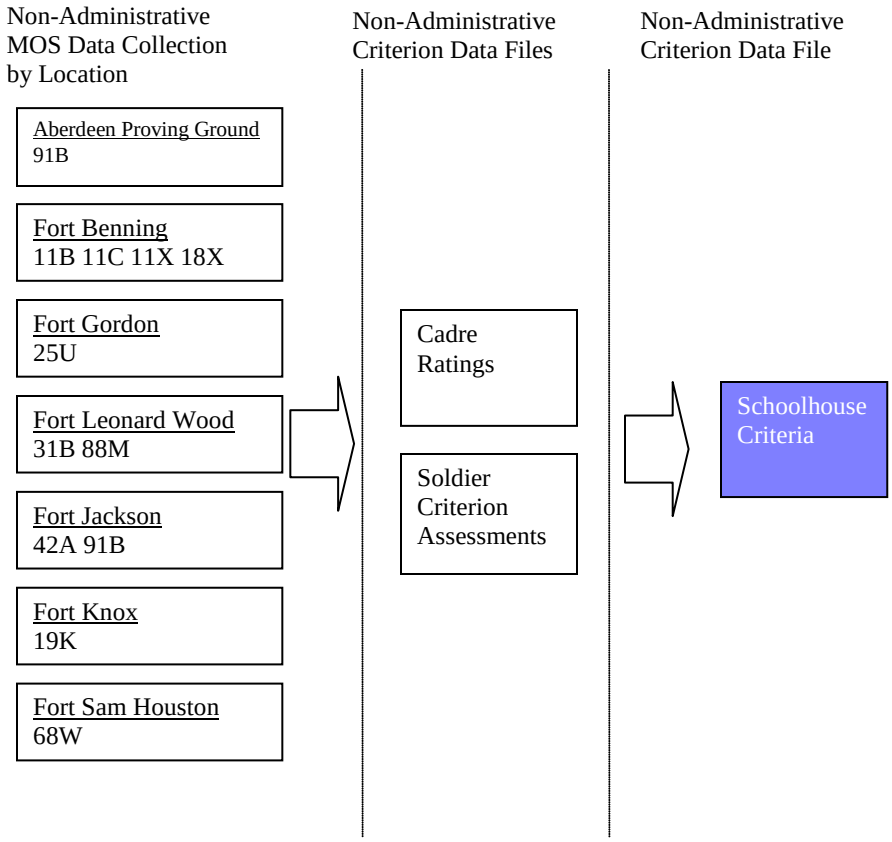

Figure 2.1. Summary of TOPS schoolhouse (IMT) data sources.

A broader view of the entire TOPS analysis file construction process is provided in Figure 2.2. The lighter boxes within the figure represent source data files, and the darker boxes represent samples on which descriptive or inferential analyses are conducted. Samples are formed by applying filters to a data file such that it includes the observations of interest. The leftmost column in the figure summarizes the predictor data sources used to derive the TOPS

[5] Administrative data are collected from the following sources: (a) Military Entrance Processing Command (MEPCOM), (b) Army Human Resources Command (AHRC), (c) U.S. Army Accessions Command (USAAC), and (d) Army Training Support Center (ATSC).

Applicant Sample. The other columns summarize the research-only (i.e., non-administrative) and administrative criterion data. Predictor and criterion data are merged to form the schoolhouse-specific validation sample and the full validation sample. The latest version of the TOPS data file does not contain WPA or ICTL predictor scores or in-unit criteria. Future versions of the data file will include those data.

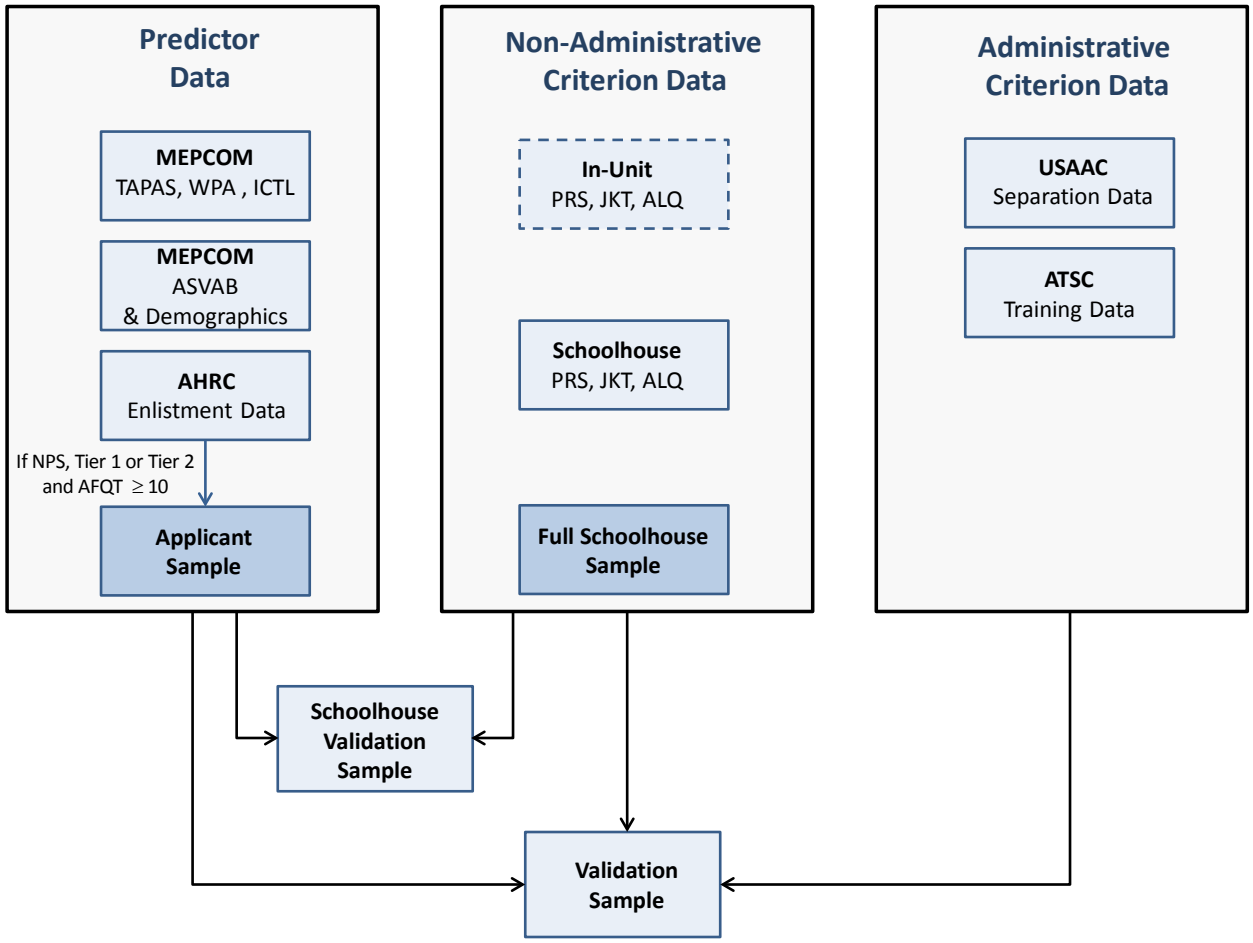

Figure 2.2. Overview of TOPS data file merging and nested sample generation process.

Description of Data File and Sample Construction

Table 2.1 summarizes the total TAPAS sample contained in the June 2011 TOPS data file by key variables that were used to create the samples on which analyses were conducted. The total sample includes applicants who did not enlist in the Army. The majority of individuals in the data file are classified as Education Tier 1 or 2, non-prior service, and AFQT Category I to IV (i.e., AFQT score ≥ 10). All analyses are restricted to these individuals, which results in elimination of approximately 7% of the total records in the data file.

Table 2.1. Full TAPAS Data File Sample Characteristics

Variables	*n*	% of Total Sample (*N* = 151,625)
Education Tier		
Tier 1	142,071	93.7
Tier 2	5,830	3.8
Tier 3	3,722	2.5
Prior Service		
Yes	4,116	2.7
No or Missing	147,509	97.3
Military Occupational Specialty		
11B/11C/11X/18X	10,955	6.9
19K	752	0.5
25U	1,113	0.7
31B	2,386	1.6
42A	1,143	0.8
68W	3,425	2.3
88M	3,037	2.0
91B	2,701	1.8
Other	36,246	24.2
Unknown[a]	89,867	59.2
AFQT Category		
I	11,740	7.7
II	47,291	31.2
IIIA	30,223	19.9
IIIB	36,989	24.4
IV[b]	22,849	15.1
V	2,528	1.7
Contract Status		
Signed	67,642	44.6
Not signed	83,983	55.4
Total Applicant Sample[c]	141,483	93.3

[a] Generally, when the MOS is unknown, it is either because the respondent did not access into the Army or because the information was not yet available in the data sources on which the June 2011 data file was based.
[b] AFQT Category IV is oversampled. Figures presented are not representative of Army accessions.
[c] The Applicant Sample size is smaller than the full TAPAS sample because it is limited to non-prior service, Education Tier 1 and 2, AFQT ≥ 10 applicants.

The number and percentage of each MOS represented in the schoolhouse criterion data file are found in Table 2.2. The MOS represented most heavily are11B and 68W; least well represented are 19K, 25U, and 42A Soldiers.

Table 2.2. Distribution of MOS in the Full Schoolhouse Data File

MOS	*n*	*%*
11B/11C/11X/18X[a]	9,959	39.7
19K	135	0.5
25U	469	1.9
31B	3,546	14.1
42A	700	2.8
68W	5,629	22.4
88M	3,230	12.9
91B	1,429	5.7
Other	18	0.1
Total	25,115	100.0

[a]At this stage, infantry includes multiple MOS designators.

A detailed breakout of background and demographic characteristics observed in the analytic samples appears in Table 2.3. Regular Army Soldiers comprise a majority of the cases in each sample. AFQT categories follow an expected distribution. The samples are predominantly male, Caucasian, and non-Hispanic; however, a significant percentage of Soldiers declined to provide information on race or ethnicity. The TOPS Applicant Sample was defined by limiting records in the full data file to those Soldiers who are non-prior service, Education Tier 1 or 2, and have an AFQT score of 10 or greater.

The Validation Sample described in Table 2.3 includes 46,188 Soldiers. Included in this sample are Soldiers who meet all of the inclusion criteria for the TOPS Applicant Sample and also have at least one record in a criterion data source (i.e., Army Training Requirements and Resources System [ATRRS], Resident Individual Training Management System [RITMS], IMT/schoolhouse, attrition). However, the number of Soldiers included in any individual analysis is generally much smaller. The exact number of Soldiers included in a given analysis depends on the criterion variable involved and the limiting factors imposed on that variable (e.g., usability flags, limitations on component). Specific sample details on each criterion variable are provided in subsequent chapters. Generally speaking, administrative graduation and exam records represent the most available criterion data source, followed by 3-month attrition, which consists of data for nearly 25,000 Soldiers.

Although there are 25,115 Soldiers in the full schoolhouse data file, only 4,976 had taken the TAPAS when they applied for enlistment. This disconnect is due largely to the fact that most of the Soldiers tested at the schools had taken their pre-enlistment tests in 2009, before MEPCOM started administering the TAPAS widely to applicants. The problem is exacerbated by the gradual introduction of the TAPAS across MEPS locations so that early in the IOT&E, not all MEPS were actively participating. Another contributing factor is the extended time, ranging from approximately 6-9 months, from when the applicants complete enlistment testing and access into the Army. We expect that future analysis data files will continue to show a higher match between Soldiers tested in the schools and those tested pre-enlistment. Indeed, the match rate at this stage (19.8%) is an improvement over the match rates obtained previously (5.5%, 12.7%; Trippe, Ford, Bynum, & Moriarty, 2011).

Table 2.3. Background and Demographic Characteristics of the TOPS Samples

	Applicant[a]		Validation[b]		Schoolhouse Validation[c]	
	n = 141,483		*n* = 46,188		*n* = 4,976	
Characteristic	*n*	*%*	*n*	*%*	*n*	*%*
Component						
Regular	87,322	61.7	28,279	61.2	3,093	62.2
ARNG	37,594	26.6	12,767	27.6	1,465	29.4
USAR	16,452	11.6	5,140	11.1	418	8.4
Military Occupational Specialty						
11B/11C/11X/18X	10,572	7.5	8,688	18.8	2,254	45.3
19K	740	0.5	570	1.2	53	1.1
25U	1,078	0.8	712	1.5	10	0.2
31B	2,280	1.6	1,733	3.8	760	15.3
42A	1,095	0.8	741	1.6	107	2.2
68W	3,328	2.4	2,613	5.7	945	19.0
88M	2,874	2.0	1,996	4.3	624	12.5
91B	2,578	1.8	1,901	4.1	223	4.5
Other	34,656	24.5	27,156	58.8	--	--
Unknown	82,282	58.2	78	0.2	--	--
AFQT Category						
I	11,059	7.8	4,075	8.8	423	8.5
II	45,051	31.8	16,778	36.3	1,977	39.7
IIIA	28,926	20.4	10,456	22.6	1,116	22.4
IIIB	35,103	24.8	12,540	27.1	1,276	25.6
IV	21,344	15.1	2,339	5.1	184	3.7
Gender						
Female	27,540	19.5	7,461	16.2	586	11.8
Male	113,875	80.5	38,726	83.8	4,390	88.2
Race						
African American	20,722	14.6	6,057	13.1	465	9.3
American Indian	1,014	0.7	313	0.7	40	0.8
Asian	4,285	3.0	1,424	3.1	127	2.6
Hawaiian/Pacific Islander	1,057	0.7	434	0.9	57	1.1
Caucasian	102,819	72.7	35,750	77.4	4,024	80.9
Multiple	599	0.4	202	0.4	26	0.5
Declined to Answer	10,987	7.8	2,008	4.3	237	4.8
Ethnicity						
Hispanic/Latino	20,802	14.7	6,279	13.6	544	10.9
Not Hispanic	109,903	77.7	38,187	82.7	4,224	84.9
Declined to Answer	10,777	7.6	1,722	3.7	208	4.2

[a] Limited to applicants who had no prior service, Education Tier 1 or 2, and AFQT ≥ 10; served as the core analysis sample .
[b] Limited to applicants who had no prior service, Education Tier 1 or 2, and AFQT ≥ 10 and had a record in one of the sources used for criterion analyses (i.e., schoolhouse, ATTRS, RITMS, or attrition).
[c] Applicants with schoolhouse data who also had a record in the full TOPS data file.

Summary

The TOPS data file is periodically updated by merging TAPAS scores, administrative records, and IMT data into one master data file. The June 2011 data file includes a total of 151,625 applicants who took the TAPAS, 141,483 of whom were in the TOPS Applicant Sample. The Applicant Sample was determined by excluding Education Tier 3, AFQT Category V, and prior service applicants from the master data file. Of that Applicant Sample, 46,188 (33%) had a record in at least one of the criterion data sources and 4,976 (3.5%) had IMT data collected from the schoolhouse. The schoolhouse match rate represents an improvement from the prior reporting cycle. This is likely due to the maturation of criterion data in the source data files. Higher match rates observed in the present reporting cycle are likely to improve the stability and interpretability of results over the prior cycle. Nevertheless, the amount of criterion data that is actually used in a given analysis remains small in relation to the amount of available predictor data. Subsequent iterations of the TOPS IOT&E data file will no doubt show progressively stronger sample sizes to support validation and other evaluative analyses.

CHAPTER 3: DESCRIPTION OF THE TOPS IOT&E PREDICTOR MEASURES

Stephen Stark, O. Sasha Chernyshenko, Fritz Drasgow (Drasgow Consulting Group), and Matthew T. Allen (HumRRO)

The purpose of this chapter is to describe the predictor measures investigated to date in the TOPS IOT&E (i.e., TAPAS and ASVAB). The central predictor under investigation in this analysis is TAPAS (Stark, Chernyshenko, & Drasgow, 2010b), while the baseline predictor used by the Army is the ASVAB. Two additional experimental measures, the ICTL and WPA, are not yet included in the analysis data files, and are therefore not discussed further here. We begin this chapter by describing the TAPAS, including previous research and scoring methodology. This is followed by a brief description of the versions administered as part of the TOPS IOT&E. We conclude with a brief description of the ASVAB.

Tailored Adaptive Personality Assessment System (TAPAS)

TAPAS Background

TAPAS is a personality measurement tool developed by Drasgow Consulting Group (DCG) under the Army's Small Business Innovation Research (SBIR) program. The system builds on the foundational work of the Assessment of Individual Motivation ([AIM]; White & Young, 1998) by incorporating features designed to promote resistance to faking and by measuring narrow personality constructs (i.e., facets) that are known to predict outcomes in work settings. Because TAPAS uses item response theory (IRT) methods to construct and score items, it can be administered in multiple formats: (a) as a fixed length, *nonadaptive test* where examinees respond to the same sequence of items or (b) as an *adaptive* test where each examinee responds to a unique sequence of items selected to maximize measurement accuracy for that specific examinee.

TAPAS uses an IRT model for multidimensional pairwise preference items ([MUPP]; Stark, Chernyshenko, & Drasgow, 2005) as the basis for constructing, administering, and scoring personality tests that are designed to reduce response distortion (i.e., faking) and yield normative scores even with tests of high dimensionality (Stark, Chernyshenko, & Drasgow 2010a). TAPAS items consist of pairs of personality statements for which a respondent's task is to choose the one that is "more like me." The two statements constituting each item are matched in terms of social desirability and often represent different dimensions. As a result, respondents have a difficult time discerning which answers improve their chances of being enlistment eligible. Because they are less likely to know which dimensions are being used for selection, they are less likely to discern which statements measure those dimensions, and they are less likely to be able to keep track of their answers on several dimensions simultaneously so as to provide consistent patterns of responses across the whole test. Without knowing which answers have an impact on their eligibility status, respondents should not be able to increase their scores on selection dimensions as easily as when traditional, single statement measures are used.

The use of a formal IRT model also greatly increases the flexibility of the assessment process. A variety of test versions can be constructed to measure personality dimensions that are

relevant to specific work contexts, and the measures can be administered via paper-and-pencil or computerized formats. If test content specifications (i.e., test blueprints) are comparable across versions, the respective scores can be readily compared because the metric of the statement parameters has already been established by calibrating response data obtained from a base or reference group (e.g., Army recruits). The same principle applies to adaptive testing, wherein each examinee receives a different set of items chosen specifically to reduce the error in his or her trait scores at points throughout the exam. Adaptive item selection enhances test security because there is less overlap across examinees in terms of the items presented. Even with constraints governing the repetition and similarity of the psychometric properties of the statements composing TAPAS items, we estimate that over 100,000 possible pairwise preference items can be crafted from the current 15-dimension TAPAS pool.

Another important feature of TAPAS is that it contains statements representing 22 narrow personality traits. The TAPAS trait taxonomy was developed using the results of several large scale factor-analytic studies with the goal of identifying a comprehensive set of non-redundant narrow traits. These narrow traits, if necessary or desired, can be combined to form either the Big Five (the most common organization scheme for narrow personality traits) or any other number of broader traits (e.g., Integrity or Positive Core Self-Evaluations). This is advantageous for applied purposes because TAPAS versions can be created to fit a wide range of applications and are not limited to a particular service branch or criterion. Selection of specific TAPAS dimensions can be guided by consulting the results of a meta-analytic study performed by DCG that mapped the 22 TAPAS dimensions to several important organizational criteria for military and civilian jobs (e.g., task proficiency, training performance, attrition) (Chernyshenko & Stark, 2007).

Three Current Versions of TAPAS

As part of the TOPS IOT&E, three versions of the TAPAS were administered. The first was a 13-dimension computerized adaptive test (CAT) containing 104 pairwise preference items. This version is referred to as the TAPAS-13D-CAT, and was administered from May 4, 2009 to July 10, 2009 to over 2,200 Army and Air Force recruits.[6] In July 2010, ARI decided to expand the TAPAS to 15 dimensions by adding the facets of Adjustment from the Emotional Stability domain and Self Control from the Conscientiousness domain. Test length was also increased to 120 items. Two 15-dimension TAPAS tests were created. One version was nonadaptive (static), so all examinees answered the same sequence of items; the other was adaptive, so each examinee answered items tailored to his or her trait level estimates. The TAPAS-15D-Static was administered from mid-July to mid-September of 2009 to all examinees, and later to smaller numbers of examinees at some MEPS. The adaptive version, referred to as TAPAS-15D-CAT, was introduced in September 2009 and Army, Air Force, and Navy recruits continue to complete this version.[7] Table 3.1 shows the facets assessed by the 13-dimension and 15-dimension measures.

[6] Note that MEPCOM also is administering the TAPAS to Air Force applicants on an experimental basis.
[7] Navy recruits began taking the TAPAS 1 April 2011.

Table 3.1. TAPAS Dimensions Assessed

Facet Name	Brief Description	"Big Five" Broad Factor
Dominance	High scoring individuals are domineering, "take charge" and are often referred to by their peers as "natural leaders."	Extraversion
Sociability	High scoring individuals tend to seek out and initiate social interactions.	
Attention Seeking	High scoring individuals tend to engage in behaviors that attract social attention; they are loud, loquacious, entertaining, and even boastful.	
Generosity	High scoring individuals are generous with their time and resources.	Agreeableness
Cooperation	High scoring individuals are trusting, cordial, non-critical, and easy to get along with.	
Achievement	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.	Conscientiousness
Order	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.	
Self Control[a]	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.	
Non-Delinquency	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.	
Adjustment[a]	High scoring individuals are worry free, and handle stress well; low scoring individuals are generally high strung, self-conscious and apprehensive.	Emotional Stability
Even Tempered	High scoring individuals tend to be calm and stable. They don't often exhibit anger, hostility, or aggression.	
Optimism	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.	
Intellectual Efficiency	High scoring individuals are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual.	Openness To Experience
Tolerance	High scoring individuals scoring are interested in other cultures and opinions that may differ from their own. They are willing to adapt to novel environments and situations.	
Physical Conditioning	High scoring individuals routinely participate in vigorous sports or exercise and enjoy physical work.	Other

[a]Not included in TAPAS-13D-CAT.

As part of the first TOPS IOT&E evaluation cycle, descriptive statistics for the TAPAS were computed along with analyses examining the equivalence of these three forms. In general, the results suggested that the three forms were equivalent, and thus could be treated as the same measure provided that the values were standardized within version (Allen, Ingerick, & DeSimone, 2011). With this in mind, the TOPS TAPAS versions were combined into one overall set of TAPAS scales by:

1. Filtering out participants who were not part of the sample of interest (i.e., those that were not in the "TOPS Applicant Sample"—Education Tier 1 and 2, non-prior service, AFQT Category IV or above), and

2. Standardizing the variables within version using a *z*-transformation, completed by subtracting each score from the mean for that version and dividing by the standard deviation.

TAPAS Scoring

TAPAS scoring is based on the MUPP IRT model originally proposed by Stark (2002). The model assumes that when person *j* encounters stimuli *s* and *t* (which, in our case, correspond to two personality statements), the person considers whether to endorse *s* and, independently, considers whether to endorse *t*. This process of independently considering the two stimuli continues until one and only one stimulus is endorsed. A preference judgment can then be represented by the joint outcome (Agree with *s*, Disagree with *t*) or (Disagree with *s*, Agree with *t*). Using a 1 to indicate agreement and a 0 to indicate disagreement, the outcome (1,0) indicates that statement *s* was endorsed but statement *t* was not, leading to the decision that *s* was preferred to statement *t*; an outcome of (0,1) similarly indicates that stimulus *t* was preferred to *s*. Thus, the probability of endorsing a stimulus *s* over a stimulus *t* can be formally written as

$$P_{(s>t)_i}(\theta_{d_s},\theta_{d_t}) = \frac{P_{st}\{1,0 \mid \theta_{d_s},\theta_{d_t}\}}{P_{st}\{1,0 \mid \theta_{d_s},\theta_{d_t}\} + P_{st}\{0,1 \mid \theta_{d_s},\theta_{d_t}\}},$$

where:

$P_{(s>t)_i}(\theta_{d_s},\theta_{d_t}) =$ probability of a respondent preferring statement *s* to statement *t* in item *i*,

$i =$ index for items (i.e., pairings), where $i = 1$ to I,

$d =$ index for dimensions, where $d = 1, \ldots, D$, d_s represents the dimension assessed by statement *s*, and d_t represents the dimension assessed by statement t,

$s, t =$ indices for first and second statements, respectively, in an item,

$(\theta_{d_s},\theta_{d_t}) =$ latent trait scores for the respondent on dimensions d_s and d_t respectively,

$P_{st}(1,0 \mid \theta_{d_s}, \theta_{d_t}) =$ joint probability of endorsing stimulus *s* and not endorsing stimulus *t* given latent trait scores $(\theta_{d_s}, \theta_{d_t})$,

and

$P_{st}(0,1 \mid \theta_{d_s}, \theta_{d_t}) =$ joint probability of not endorsing stimulus *s* and endorsing stimulus *t* given latent trait scores $(\theta_{d_s}, \theta_{d_t})$.

With the assumption that the two statements are evaluated independently, and with the usual IRT assumption that only θ_{d_s} influences responses to statements on dimension d_s and only θ_{d_t} influences responses to dimension d_t (i.e., local independence), we have

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t})}{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t}) + P_s(0 \mid \theta_{d_s}) P_t(1 \mid \theta_{d_t})},$$

where

$P_s(1 \mid \theta_{d_s}), P_s(0 \mid \theta_{d_s}) =$ probability of endorsing/not endorsing stimulus *s* given the latent trait value θ_{d_s},

and

$P_t(0 \mid \theta_{d_t}), P_t(1 \mid \theta_{d_t}) =$ probability of endorsing/not endorsing stimulus *t* given latent trait θ_{d_t} .The probability of preferring a particular statement in a pair thus depends on θ_{d_s} and θ_{d_t}, as well as the model chosen to characterize the process for responding to the individual statements. Toward that end, Stark (2002) proposed using the dichotomous case of the generalized graded unfolding model ([GGUM]; Roberts, Donoghue, & Laughlin, 2000), which has been shown to fit personality data reasonably well (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006).

Test scoring is done via Bayes modal estimation. For a vector of latent trait values, $\tilde{\theta} = (\theta_{d'=1}, \theta_{d'=2}, ..., \theta_{d'=D})$, this involves maximizing:

$$L(\tilde{u}, \tilde{\theta}) = \left\{ \prod_{i=1}^{n} \left[P_{(s>t)_i} \right]^{u_i} \left[1 - P_{(s>t)_i} \right]^{1-u_i} \right\} * f(\tilde{\theta}) \quad ,$$

where $\tilde{u}$ is a binary response pattern, $P_{(s>t)}$ is the probability of preferring statement *s* to statement *t* in item *i*, and $f(\tilde{\theta})$ is a *D*-dimensional prior density distribution, which, for simplicity, is assumed to be the product of independent normals, $\prod_{d'=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\theta_{d'}^2}{2\sigma^2}}$.

Taking the natural log, for convenience, the above equation can be rewritten as:

$$\ln L(\tilde{u},\tilde{\theta}) = \sum_{i=1}^{n}\left[(u_i)\ln P_{(s>t)_i} + (1-u_i)\ln(1-P_{(s>t)_i})\right] + \sum_{d'=1}^{D}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\theta_{d'}^2}{2\sigma^2}\right] \quad ,$$

leaving the following set of equations to be solved numerically:

$$\frac{\partial \ln L}{\partial \tilde{\theta}} = \begin{bmatrix} \frac{\partial \ln L}{\partial \theta_{d'=1}} \\ \frac{\partial \ln L}{\partial \theta_{d'=2}} \\ \vdots \\ \frac{\partial \ln L}{\partial \theta_{d'=D}} \end{bmatrix} = 0$$

This equation can be solved numerically to obtain a vector of trait score estimates for each respondent via a *D*-dimensional maximization procedure (e.g., Press, Flannery, Teukolsky, & Vetterling, 1990), involving the posterior and its first derivatives. Standard errors for TAPAS trait scores are estimated using a replication method developed by Stark and colleagues (2010a). In brief, this method involves using the IRT parameter estimates for the items that were administered to generate 30 new response patterns based on an examinee's TAPAS trait scores. The resulting simulated response patterns are then scored and the standard deviations of the respective trait estimates over the 30 replications are used as standard errors for the original TAPAS values. In a recent simulation study (Stark, Chernyshenko, & Drasgow, 2010c), this new replication method provided standard error estimates that were much closer to the empirical (true) standard deviations than previously used approaches (i.e., based on the approximated inverse Hessian matrix or a jack-knife approach).

In the present research, TAPAS data were flagged as unusable if the applicant selected the same response option more than 63% of the time or, alternatively, if the applicant responded to more than two items in less than 2 seconds each. Descriptive statistics, subgroup differences, and scale intercorrelations for the TAPAS scale scores in the current sample can be found in Appendix A.

TAPAS Initial Validation Effort

Initial predictive and construct-related validity evidence on the TAPAS was collected during ARI's *Expanded Enlistment Eligibility Metrics (EEEM)* research project in 2007-2009 (Knapp & Heffner, 2010). In the EEEM project, new Soldiers completed a 12-dimension, 95-item nonadaptive (or static) version of TAPAS, called TAPAS-95s. The TAPAS-95s showed evidence of construct and criterion validity. Intellectual Efficiency and Curiosity, for example, showed moderate positive correlations with AFQT and correlations of .35 with each other. This was expected, given that both facets tap the intellectance aspects of the Big Five factor, Openness to Experience. The same two traits exhibited similarly positive, but smaller correlations with Tolerance, another facet of Openness reflecting comfortableness around others

having different customs, values, or beliefs (Chernyshenko, Stark, Woo, & Conz, 2008). TAPAS-95s dimensions also showed incremental validity over AFQT in predicting several performance criteria. For example, when TAPAS trait scores were added to the regression analysis based on a sample of several hundred Soldiers, the multiple correlation increased by .35 for the prediction of physical fitness, .20 for the prediction of disciplinary incidents, and .11 for the prediction of 6-month attrition. None of these criteria were predicted well by AFQT alone (predictive validity estimates were consistently below .10).

The first TOPS IOT&E report expanded on these results by comparing the psychometric properties of the TOPS TAPAS and TAPAS-95s (Knapp, Heffner, et al., 2011). The results of these analyses suggested that (a) the standard deviations for the TOPS TAPAS were, on average, smaller than those found for the TAPAS-95s; (b) some TAPAS scales were more similar across the two settings than others (e.g., Physical Conditioning was consistent, while Attention Seeking was not); and (c) the TOPS TAPAS scales were not strongly related to other individual difference variables (e.g., race, gender), consistent with what was found in EEEM (Allen et al., 2011).

The validity of the TAPAS for predicting key performance and retention-related outcomes of interest in an applicant environment has been examined in the last two TOPS IOT&E technical reports (Knapp & Heffner, 2011; Knapp, Heffner, et al., 2011). While the sample sizes for the initial validation analyses were too small to yield stable estimates, the second set of analyses conducted for the TOPS 2010 summary report were much more revealing (Caramagno, Allen, & Ingerick, 2011). Multiple TAPAS scales, such as Dominance, Physical Conditioning, and Optimism, consistently predicted key criteria of interest. Additionally, the effects for will-do performance and retention-related criteria were largely independent of AFQT. However, the initial TOPS composites (described below) only demonstrated partial utility for identifying low potential candidates to "select out." These results suggest that the composites should be reconceptualized to better account for changes from the experimental to applicant settings, a point that is addressed more fully in Chapter 5.

Initial TAPAS Composites

In addition to the validation analyses described above, an initial Education Tier 1 performance screen was developed from the TAPAS-95s scales for the purpose of testing in an applicant setting (Allen et al., 2010). This was accomplished by (a) identifying key criteria of most interest to the Army, (b) sorting these criteria into "can-do" and "will-do" categories (see below), and (c) selecting composite scales corresponding to the can-do and will-do criteria, taking into account both theoretical rationale and empirical results. The result of this process was two composite scores.

1. Can-Do Composite: The TOPS can-do composite consists of five TAPAS scales and is designed to predict can-do criteria, such as MOS-specific job knowledge, Advanced Individual Training (AIT) exam grades, and graduation from AIT/One Station Unit Training (OSUT).

2. Will-Do Composite: The TOPS will-do composite consists of five TAPAS scales (three of which overlap with the can-do composite) and is designed to predict will-do criteria such as physical fitness, adjustment to Army life, effort, and support for peers.

The analyses on which these composites were based focused on Tier 1 AFQT Category IIIB applicants. Due to changing recruitment priorities (as described in Chapter 1), the initial target group for the TOPS IOT&E was AFQT Category IV applicants, who must score above the 10th percentile on both the can-do and will-do TAPAS. Subsequently, the TOPS IOT&E was expanded to include all Tier 1 and Tier 2 applicants above AFQT Category V, but screening based on TAPAS scores is confined to Category IIIB and IV Tier 1 applicants.

Armed Services Vocational Aptitude Battery (ASVAB) Content, Structure, and Scoring

The ASVAB is a multiple aptitude battery of nine tests administered by the MEPCOM. Most military applicants take the computer adaptive version of ASVAB (i.e., the CAT-ASVAB). Scores on the ASVAB tests are combined to create composite scores for use in (a) selecting applicants into the Army and (b) classifying them to an MOS. The AFQT comprises the Verbal Expression[8] (VE), Arithmetic Reasoning (AR), and Math Knowledge (MK) tests (AFQT = 2*VE + AR + MK). Applicants must meet a minimum AFQT score to be eligible to serve in the military, and the Services favor high-scoring applicants for enlistment (e.g., through enlistment bonuses). AFQT percentile scores are divided into the following categories:[9]

- Category I (93-99)
- Category II (65-92)
- Category IIIA (50-64)
- Category IIIB (31-49)
- Category IV (10-30)[10]
- Category V (1-9)

AFQT Category V Soldiers are not eligible for enlistment, while no more than 20% of the total number of enlisted Soldiers can be AFQT Category IV. AFQT Category IIIB applicants are also given lower enlistment priority than AFQT Category I to IIIA applicants.

For classification, scores on the ASVAB tests are combined to form nine Aptitude Area (AA) composites.[11] An applicant must receive a minimum score on the MOS-relevant AA composite(s) to qualify for classification to that MOS. For example, applicants must score a 95 in both the Electronics (EL) and Signal Communications (SC) AA composites to qualify as a Signal Support Specialist (25U). Descriptive statistics for the AFQT, ASVAB tests, and AA

[8] Verbal Expression is a scaled combination of the Word Knowledge (WK) and Paragraph Comprehension (PC) tests.
[9] For more information on ASVAB scoring, see the official website of the ASVAB, www.officialasvab.com.
[10] AFQT Category IV can be further subdivided into IVA (21-30), IVB (16-20), and IVC (15-15). For the purposes of this report, all AFQT Category IV Soldiers are treated as one group.
[11] A tenth AA composite, General Technical (GT), is not used for enlisted Army selection or classification and, therefore, is not included here.

composites are reported in Table A.3 in Appendix A. AFQT Category frequencies are reported in Chapter 2 (Tables 2.1 and 2.3).

Summary

The purpose of this chapter was to describe the predictor measures used as part of the TOPS IOT&E. Three versions of the experimental measure—the TAPAS—were administered as part of the TOPS IOT&E. The TAPAS is unique among typical personality measures because it uses forced-choice pairwise items and IRT to promote resistance to faking. Initial validation research conducted as part of EEEM was promising enough to warrant an IOT&E. Comparative analyses suggest that the three versions of the TAPAS are equivalent, but indicated some differences with the TAPAS-95s administered as part of EEEM. The ASVAB will be used as the baseline instrument for these analyses, which consists of multiple tests that are formed into selection (i.e., AFQT) and classification (i.e., AA) composites.

CHAPTER 4: DESCRIPTION AND PSYCHOMETRIC PROPERTIES OF CRITERION MEASURES

Karen O. Moriarty and Tina Chang (HumRRO)

Training criterion measures such as job knowledge tests (JKTs), performance rating scales (PRS), and attitudinal data captured on a self-report questionnaire were used to validate the TAPAS. These measures were originally developed for the training phase of the Army Class project (Moriarty, Campbell, Heffner, & Knapp, 2009), and modified, where needed, for inclusion in the TOPS IOT&E. These measures were used to supplement the administrative data. Table 4.1 summarizes the training criterion measures.

Table 4.1. Summary of Training Criterion Measures

Criterion Measure	Description
Soldier/Cadre Reported	
Job Knowledge Tests (JKT)	MOS-specific JKTs measure Soldiers' knowledge of basic facts, principles, and procedures required of Soldiers in training for a particular MOS. Each JKT includes a mix of item formats (e.g., multiple-choice, multiple-response, and rank order). The Warrior Tasks and Battle Drills (WTBD) JKT measures knowledge that is general to all enlisted Soldiers.
Performance Rating Scales (PRS)	PRS measure Soldiers' training performance on two categories: (a) MOS-specific (e.g., learns preventive maintenance checks and services, learns to troubleshoot vehicle and equipment problems) and (b) Army-wide (e.g., exhibits effort, supports peers, demonstrates physical fitness). The PRS are completed by drill sergeants or training cadre.
Army Life Questionnaire (ALQ)	ALQ measures Soldiers' self-reported attitudes and experiences through IMT. The training ALQ focuses on Soldiers' attitudes and experiences in IMT and includes 13 scales that cover (a) Soldiers' commitment and retention-related attitudes and (b) Soldiers' performance and adjustment.
Administrative	
Attrition	Attrition data were obtained on participating Regular Army Soldiers at 3, 6, and 9 months time in service (TIS).
Initial Military Training (IMT) Criteria	These data provide information concerning how many Soldiers restarted IMT and for what reasons, the number of times Soldiers restarted training, and graduation status.
AIT School Grades	Schoolhouse grades for Soldiers in Advanced Individual Training (AIT).

Training Criterion Measure Descriptions

Job Knowledge Tests (JKTs)

Seven JKTs were developed or adapted for this research: one for Warrior Tasks and Battle Drills (WTBD), which is common for all Soldiers, and six MOS-specific JKTs for Infantry, Armor, Military Police, Health Care Specialist, Light Wheel Vehicle Mechanic, and Motor Transport Operator. Depending upon the MOS, many JKT items were drawn from items originally developed in prior ARI projects (Campbell & Knapp, 2001; Collins, Le, & Schantz, 2005; Knapp & Campbell, 2006). Most of the JKT items are in a multiple-choice format with two to four response options. However, other formats, such as multiple response (i.e., check all that apply), rank ordering, and matching are also used. The items use visual images to make them more realistic and reduce reading requirements for the test.

Prior to finalizing the items for use in the TOPS IOT&E, the items were reviewed to ensure they were of high quality. First, we examined the comments Soldiers provided about the assessments during the Army Class testing sessions and made corrections where necessary. For example, several 11B Soldiers did not know the meaning of the word, "demarcate," so we changed that word to "mark." Second, we reviewed item statistics from the Army Class data and dropped items that had poor item statistics (e.g., low item-total correlations). Finally, results of the Army Class JKT analyses suggested that the training JKTs were too difficult, so we eliminated the more difficult items.

Performance Rating Scales (PRS)

The PRS also have roots in previous research (see Moriarty et al., 2009 for details). Table 4.2 provides descriptions of two example scales. Depending on MOS, the number of dimensions ranges from five to nine. The scales were completed by cadre members (supervisors/trainers) of the target Soldiers. The scales range from 1 (lowest) to 7 (highest) and include a "not observed" option for instances where the cadre did not have an opportunity to observe a Soldier's performance on a particular dimension. They are in the format of a behaviorally-anchored rating scale (BARS). In a BARS format, raters provide one rating for each dimension of performance. To assist in their ratings, the scales include several examples (called "anchors") of high, medium, and low performance. Figure 4.1 provides an example of one of the BARS administered.

Table 4.2. Example Training Performance Rating Scales

MOS/Army-Wide	Scale Name	Description
Army-Wide	Effort	Puts forth individual effort in study, practice, preparation, and participation activities to complete AIT/OSUT requirements to meet individual Soldier expectations.
MOS-Specific	Area Security	How well has the Soldier learned to function as a member of a lead or trail team while providing security for a convoy in a tactical environment?

Effort Puts forth individual effort in study, practice, preparation, and participation activities to complete AIT/OSUT requirements and to meet individual Soldier expectations.		
1 2	3 4 5	6 7
– Puts off studying and practicing tasks. – May tune out while an instructor is speaking and sometimes isn't prepared for class. – Tends to give up on tasks if problems arise.	– Usually completes required assignments. – Pays attention in class and is usually adequately prepared for class. – Usually keeps trying when problems arise.	– Completes study and practice assignments, including non-class requirements, on time. – Pays attention in class and studies hard in preparation for class. – Persists with tasks even when problems arise.

Figure 4.1. Sample 7-point behaviorally-anchored rating scale.

In addition to the BARS ratings of each performance dimension, respondents were also asked to provide one rating assessing overall performance. This rating was made on a 5-point relative comparison scale, as shown in Figure 4.2. The PRS assessment also includes an initial 3-point "familiarity" rating in which the rater indicates his or her general opportunity to observe each Soldier being rated (i.e., limited, reasonable, or a lot of opportunity to observe).

Prior IOT&E evaluations noted low inter-rater reliability estimates for the PRS,[12] so steps are underway to change the format of the rating scales in an effort to improve their psychometric characteristics. Specifically, both the Army-wide and MOS-specific PRS dimension ratings will be converted to a relative scale that parallels the format of the overall rating scale shown in Figure 4.2. We believe that this format is more suitable for a training environment in which raters observe a very large number of Soldiers for a relatively short period of time. The familiarity scale will also be changed to a 4-point scale in which raters can more clearly indicate their ability to judge each Soldier's performance. The newly formatted rating scales were introduced into the data collection effort in fall 2011 and are thus not included in the data file that was analyzed for the present report.

Overall Performance Considering your evaluation of the Soldier on the dimensions important to successful performance, please rate the overall effectiveness of each Soldier compared to his/her peers.				
1	2	3	4	5
Among the Weakest	**Below Average**	**Average**	**Above Average**	**Among the Best**
(in the bottom 20% of Soldiers)	(in the bottom 40% of Soldiers)	(better than the bottom 40% of Soldiers, but not as good as the top 40%)	(in the top 40% of Solders)	(in the top 20% of Soldiers)

Figure 4.2. Relative overall performance rating scale.

[12] Interrater reliability was assessed using $G(q,k)$, a reliability metric designed specifically for studies like TOPS where the measurement design is ill-structured (Putka, Le, Ingerick, & Diaz, 2008).

Army Life Questionnaire (ALQ)

The ALQ was designed to measure Soldiers' self-reported attitudes and experiences in training. An earlier form of the ALQ (Van Iddekinge, Putka, & Sager, 2005) was modified slightly for use in the TOPS IOT&E. It focuses on first-term Soldiers' attitudes and experiences in IMT and includes 13 scales that cover (a) Soldiers' commitment and retention-related attitudes and (b) Soldiers' performance and adjustment. Each ALQ scale is scored differently depending on the nature of the attribute being measured. The Army Physical Fitness Test (APFT) score is a write-in item. Training Achievements, Training Failures, and Disciplinary Incidents are simply a sum of the 'YES' responses. The remaining scales (see Table 4.3) are scored with Likert-type scales by computing a mean of the constituent item scores. To simplify the analyses, four scales administered with the ALQ and included in previous reports (Knapp & Heffner, 2011; Knapp, Heffner, et al., 2011)—Normative Commitment, Army Career Intentions, Army Reenlistment Intentions, and Army Civilian Comparison—were excluded from the current analyses.

Administrative Criteria

Attrition is a broad category that encompasses involuntary and voluntary separations for a variety of reasons (e.g., underage enlistment, conduct, family concerns, drugs or alcohol, performance, physical standards or weight, mental disorder, or violations of the Uniform Code of Military Justice [UCMJ]). The reason for separation was determined by the Soldier's Separation Program Designator (SPD) code. Soldiers who were classified as "attrits" for reasons outside of the Soldiers' or the Army's control were excluded in our analyses (e.g., death or serious injury incurred while performing one's duties).

Data on IMT school performance and completion were extracted from ATRRS and RITMS data files (see Chapter 2). ATRRS course information was used to determine (a) whether a Soldier graduated from or was discharged during IMT and (b) whether he or she restarted during IMT. RITMS data were used to determine Soldiers' AIT course grades. Given that each course has different grading procedures, the AIT course grade analysis variable was created by standardizing the grades within course. Due to restricted variance in the OSUT grades (i.e., all of the grades were pass/fail), these courses were excluded from the data file.

Training Criterion Measure Scores and Associated Psychometric Properties

Basic descriptive statistics are available for the Full Schoolhouse Sample (n = 25,115) and by MOS in Appendix B, along with the intercorrelations. In this section we review the psychometric characteristics of the criterion measures estimated using only data on those Soldiers from the TOPS Applicant Sample (i.e., Education Tier 1 and 2, non-prior service, AFQT Category IV or above) whose data were used in the criterion-related validity analyses reported in Chapter 5. This is referred to as the TOPS Validation Sample in Figure 2.2. Note that the means, standard deviations, and reliability estimates are generally similar to those for the Full Schoolhouse Sample.

Job Knowledge Tests (JKTs)

JKT records were flagged as not useable if the Soldier omitted more than 10% of the assessment items, took fewer than 5 minutes to complete the entire assessment, or chose an implausible response to one of the careless responding items.[13]

Table 4.3. ALQ Likert-Type Scales

Scale Name	Description	Number of Items	Example Item	Likert Scale Anchors
Affective Commitment	Measures Soldiers' emotional attachment to the Army.	7	I feel like I am part of the Army 'family.'	1 (strongly disagree) to 5 (strongly agree)
Normative Commitment	Measures Soldiers' feelings of obligation toward staying in the Army until the end of their current term of service.	5	I would feel guilty if I left the Army before the end of my current term of service.	1 (strongly disagree) to 5 (strongly agree)
Career Intentions	Measures Soldiers' intentions to re-enlist and to make the Army a career.	3	How likely is it that you will make the Army a career?	Varies by item: 1 (strongly disagree) to 5 (strongly agree); 1 (not at all confident) to 5 (extremely confident); 1 (extremely unlikely to 5 (extremely likely)
Reenlistment Intentions	Measures Soldiers' intention to reenlist in the Army.	4	How likely is it that you will leave the Army after completing your current term of service?	1 (strongly disagree) to 5 (strongly agree)
Attrition Cognition	Measures the degree to which Soldiers think about attriting before the end of their first term.	4	How likely is it that you will complete your current term of service?	Varies by item: 1 (strongly disagree) to 5 (strongly agree); 1 (never) to 5 (very often)
Army Life Adjustment	Measures Soldiers' transition from civilian to Army life.	9	Looking back, I was not prepared for the challenges of training in the Army.	1 (strongly disagree) to 5 (strongly agree)
Army Civilian Comparison	Measures Soldiers' impressions of how Army life compares to civilian life.	6	Indicate how you believe conditions in the Army compare to conditions in a civilian job with regards to pay.	1 (much better in the Army) to 5 (much better in civilian life)
MOS Fit	Measures Soldiers' perceived fit with their MOS.	9	My MOS provides the right amount of challenge for me.	1 (strongly disagree) to 5 (strongly agree)
Army Fit	Measures Soldiers' perceived fit with the Army.	8	The Army is a good match for me.	1 (strongly disagree) to 5 (strongly agree)

[13] The 5-minute criterion was established during the first in-unit phase of the Army Class project, which employs highly similar assessments administered via the same platform. See Knapp, Owens, et al. (2011) for details.

A single, overall raw score was computed for each JKT by summing the total number of points Soldiers earned across the JKT items. All of the multiple-choice items were worth one point. Depending on the format of the non-traditional items (e.g., multiple response), they were worth one or more points. To facilitate comparisons across MOS, we computed a percent correct score based on the maximum number of points that could be obtained on each MOS test. For the criterion-related validity analyses, we converted the total raw score to a standardized score (or *z*-score) by standardizing the scores *within* each MOS.

Table 4.4 shows the percent correct scores, as well as internal consistency reliability estimates for the six MOS-specific and the WTBD JKTs. The mean percent correct score across all six MOS-specific tests was 65.36%, with the 19K and 91B tests being the most difficult (means of 60.81% and 57.58%, respectively). Internal consistency reliability estimates were acceptable, though the WTBD JKT estimate was on the low side (.66), which is not surprising since it covers a broad range of tasks. Table 4.4 shows the correlations between the various MOS JKT scores and the WTBD JKT score. These are only moderate in size and considerably less than 1.00, suggesting that the MOS-specific JKTs and the WTBD JKT each provide some unique performance information.

Table 4.4. Descriptive Statistics and Reliability Estimates for Training Job Knowledge Tests (JKTs) in the Validation Sample

Test Scores	*n*	*M*	*SD*	*Min*	*Max*	r_{WTBD}	α
11B/11C/11X/18X	1,731	62.23	10.13	27.91	86.05	.57	.77
19K	47	60.81	10.55	30.00	78.00	.59	.76
31B	672	70.68	8.17	45.63	91.26	.48	.76
68W	849	74.52	10.12	33.70	92.39	.54	.86
88M	490	66.35	11.46	33.33	88.89	.63	.79
91B	148	57.58	13.90	29.90	85.57	.57	.91
WTBD Job Knowledge	4,696	66.18	12.90	9.68	96.77	--	.66

Note. Mean represents percent correct; α = coefficient alpha. WTBD = Warrior Tasks and Battle Drills. Sample = non-prior service, Education Tier 1 and 2, AFQT Category IV or above Soldiers. r_{WTBD} = correlation with WTBD JKT, all correlations are statistically significant ($p < .05$).

Performance Rating Scales (PRS)

A Soldier's PRS ratings were removed if the cadre member provided a familiarity rating of 1 ("I have had little opportunity to observe this Soldier"). PRS data also were flagged as unusable if the cadre member omitted more than 10% of the assessment items or indicated that he or she had "not observed" the individual on more than 50% of the dimensions. Data also were removed if a rater engaged in "flat responding"—that is, ratings were removed from the data file if a rater rated 10 or more Soldiers on a particular scale and 90% or more of those rating profiles were exactly the same. Approximately 20% of Soldiers with ratings data in the present sample were rated by more than one cadre member.

For the MOS-specific PRS, a composite score was created across all of the dimension scores. Creating these scores involved computing (a) the average of multiple ratings provided by the cadre (if more than one person rated the target Soldier) and (b) the mean of the average ratings on the individual scales that constitute the elements of a particular dimension. Consistent with

performance models used in previous Army research (Ingerick, Diaz, & Putka, 2009), a subset of the Army-wide scales were also combined into three unit-weighted composites: (a) Effort and Discipline PRS (a composite of the Effort and Personal Discipline scales), (b) Work with Others PRS (a composite of the Support for Peers and Peer Leadership scales), and (c) Can-Do PRS (a composite of the MOS Qualification/Skill and Common Tasks/Warrior Task Knowledge/Skill scales). Other existing scales that were not formed into composites (i.e., Physical Fitness and Bearing, Commitment and Adjustment to the Army, and Overall Performance) were included in the analysis. This composite score structure matches well with the Army-wide (AW) PRS that started being administered to Soldiers in fall 2011. We also hoped that the composite scores might show slightly better inter-rater reliability than individual scales.

Descriptive statistics and estimates of internal consistency reliability for the Army-wide and MOS-specific PRS composite scores are shown in Table 4.5. The high coefficient alphas for the Army-wide composites suggest the scale combinations were appropriate. Mean ratings are all above the mid-point, a common finding in research involving performance ratings. The ratings are also highly intercorrelated (see Appendix B, Table B.4).

Table 4.5. Descriptive Statistics and Reliability Estimates for Training Performance Rating Scales (PRS) in the Validation Sample

	n	*M*	*SD*	*Min*	*Max*	*α*	*IRR*
Army-Wide Performance Rating Scales							
Can-Do[a]	1,690	5.07	1.02	1.00	7.00	.89	.09
Commitment and Adjustment	1,705	5.15	1.13	1.00	7.00	--	.17
Effort and Discipline[a]	1,708	5.08	1.08	1.00	7.00	.83	.21
Physical Fitness and Bearing	1,700	5.01	1.13	1.00	7.00	--	.20
Work with Others[a]	1,697	5.02	1.08	1.00	7.00	.83	.17
Overall Performance	1,695	3.55	0.80	1.00	5.00	--	.36
MOS-Specific Performance Rating Composite Scores							
Total (combined across MOS)	1,425	4.93	0.93	1.00	7.00	--	--
11B/11C/11X/18X	677	5.09	0.92	1.00	7.00	.94	.25
19K	27	5.15	0.62	3.29	6.86	.88	--
31B	292	5.10	0.90	2.13	7.00	.95	.15
68W	279	4.47	0.72	1.00	7.00	.92	--
88M	126	4.80	0.96	2.00	7.00	.95	--
91B	24	4.34	1.59	1.00	7.00	.97	--

[a]Composite Army-Wide PRS comprising two dimensions. α = coefficient alpha.
Note. Sample = non-prior service, Education Tier 1 and 2, AFQT Category IV or above Soldiers. The possible Performance Rating Scale (PRS) scores are between 1 and 7, except for the Overall Performance Scale, which ranges from 1 to 5. PRS ratings were removed if the cadre member provided a familiarity rating of 1 ("I have had little opportunity to observe this Soldier"). IRR = Interrater Reliability computed using G(q,k) (Putka, Le, McCloy, & Diaz, 2008). Interrater reliability estimates are excluded if 30 or fewer Soldiers were rated by more than supervisor.

As illustrated in Table 4.5 and Appendix B, Table B.1, the interrater reliability estimates are quite low. The estimates range from .09 to .36 for the Army-wide scales in the full sample. The highest interrater reliability is associated with the Overall Performance scale on the Army-wide PRS. The low estimates on the MOS-specific rating scales are particularly disturbing given these are composite scores. We attribute these low coefficients to a few interrelated issues. First, the

number of ratees per rater is rather high, averaging 13.97 for the Full Schoolhouse Sample, which may cause the raters to become fatigued during the task. Second, most raters had very little variance in their ratings, perhaps reflecting their lack of familiarity with individual Soldiers. For example, the average within-rater standard deviation for ratings of Effort and Discipline was 0.65 in the Full Schoolhouse Sample. Third, unlike prior research (e.g., Knapp & Heffner, 2009, 2010), these data collections were not proctored. Finally, the number of raters per ratee was small, averaging less than two, which reduces the magnitude of *k*-rater interrater reliability coefficients reported in Tables 4.5 and B.1. Although not all of these potential issues with the PRS can be addressed within the practical constraints of the research (e.g., collecting ratings in an unproctored setting), the interrater reliability may be improved by the PRS format changes which were introduced in fall 2011

In Table B.4 (Appendix B), we see that the correlations among the MOS-specific PRS and the Army-wide PRS in the Full Schoolhouse Sample are moderate to large, with all of them reaching significance. These results suggest there is more content overlap between the MOS-specific PRS and the Army-wide PRS than between the MOS-specific JKTs and WTBD JKT. The Army-wide scale that correlates the strongest with the MOS-specific PRS is the Can-Do PRS composite followed by the Commitment and Adjustment scale and the Work with Others PRS composite.

Army Life Questionnaire (ALQ)

ALQ data were flagged as unusable if the Soldier omitted more than 10% of the assessment items, took fewer than 5 minutes to complete the entire assessment, or chose an implausible response to the careless responding item. In most cases, ALQ subscale scores were computed by taking the mean of all responses associated with each scale, properly accounting for reverse coded items. The Training Failures, Training Achievement, and Disciplinary Action scales were computed by summing the total number of "yes" responses.

Table 4.6 provides descriptive statistics and internal consistency reliability estimates for the training ALQ scores. Refer to Table 4.3 for scale anchors, number of items per scale, and sample items. The reliability estimates are good, ranging from .79 to .93. Mean scores are generally similar across MOS (see Table B.2 in Appendix B).

Table 4.6. Descriptive Statistics and Reliability Estimates for the Army Life Questionnaire (ALQ) in the TOPS Validation Sample

Measure/Scale	*n*	*M*	*SD*	*Min*	*Max*	*α*
Affective Commitment	4,840	3.83	0.68	1.00	5.00	.86
Attrition Cognition	4,840	1.54	0.61	1.00	5.00	.79
Army Life Adjustment	4,840	4.06	0.66	1.00	5.00	.87
MOS Fit	4,840	3.79	0.83	1.00	5.00	.93
Army Fit	4,840	4.04	0.60	1.00	5.00	.86
Training Achievement	4,830	0.40	0.61	0.00	2.00	--
Training Restarts	4,840	0.35	0.59	0.00	4.00	--
Disciplinary Incidents	3,137	0.25	0.61	0.00	6.00	--
Last APFT Score	4,785	250.51	30.77	66.00	300.00	--

Note. α = coefficient alpha. Sample = non-prior service, Education Tier 1 and 2, AFQT Category IV or above Soldiers.

Administrative Criterion Data

For the criterion variable "Restarted at Least Once During IMT," Soldiers who restarted at least once during BCT or AIT/OSUT were coded with a 0. Soldiers who completed IMT without restarting were coded with a 1. Soldiers who had not had an opportunity to fully complete their IMT at the time the data were collected were excluded from our analyses. "AIT School Grades" were computed by taking the mean of each individual's AIT course grade of record. Courses with data from fewer than 15 Soldiers were omitted from further analysis. The standardized version of this variable was created by standardizing the raw mean scores within MOS and eliminating outliers.

Table 4.7 shows descriptive statistics for the administrative variables. For Soldiers for whom data are available, the attrition rate was 6.0% for 3-month attrition, 9.4% for 6-month attrition, and 11.1% for 9-month attrition. Additionally, 13.6% of the Soldiers restarted at least once during IMT. However, it is important to note that the IMT data retrieved from administrative sources are not mature for many Soldiers. For example, although there were approximately 67,000 accessed Soldiers in the sample (see Table 2.1), we retrieved attrition data on fewer than 10,000 Soldiers and restart data on fewer than 20,000 Soldiers. Table B.8 displays the attrition and restart rates by MOS.

Table 4.7. Descriptive Statistics for Administrative Criteria in the TOPS Validation Sample

Administrative Criterion	N [a]	N_{Attrit}	$\%_{Attrit}$
Attrition			
3-Month Cumulative	24,737	1,485	6.0
6-Month Cumulative	18,917	1,779	9.4
9-Month Cumulative	11,306	1,250	11.1
Initial Military Training (IMT) Criteria	N [b]	$N_{Restarted}$	$\%_{Restarted}$
Restarted at Least Once During IMT	17,512	2,381	13.6
Restarted at Least Once During IMT for Pejorative Reasons	17,149	2,013	11.7
Restarted at Least Once During IMT for Academic Reasons	16,905	1,774	10.5
AIT School Grades	N [c]	M	SD
Overall Average (Unstandardized)	7,775	91.14	9.46
Overall Average (Standardized within MOS)	7,708	0.05	0.84

Note. Sample = non-prior service, Education Tier 1 and 2, AFQT Category IV or above Soldiers.
[a] N = number of Regular Army Soldiers with attrition data at the time data were extracted. N_{Attrit} = number of Soldiers who attrited through 3, 6, or 9 months of service. $\%_{Attrit}$ = percentage of Soldiers who attrited through 3, 6, or 9 months of service [(N_{Attrit} /N) x 100].
[b] N = number of Soldiers with IMT data at the time data were extracted. $N_{Restarted}$ = number of Soldiers who restarted at least once during IMT. $\%_{Restarted}$ = percentage of Soldiers who restarted at least once during IMT [($N_{Restarted}$ /N) x 100].
[c] N = number of Soldiers with AIT school grade data. Standardized school grades were not computed for MOS with insufficient sample size (n < 15).

Summary

Three types of measures were adapted from previous Army research to validate the TAPAS: (a) JKTs, (b) PRS, and (c) the ALQ. Additional criterion data, such as attrition, training restarts, and AIT course grades were gathered from administrative records. The JKTs are completed by Soldiers in eight target MOS and measure MOS-specific and WTBD declarative and procedural knowledge. The PRS are completed by cadre and measure MOS-specific competence and Army-wide constructs such as effort and leadership. Finally, the ALQ asks Soldiers to complete self-report verifiable performance items (e.g., their APFT scores) and attitudinal items (e.g., Adjustment to Army life). In general, the criterion measures exhibited acceptable and theoretically consistent psychometric properties. The exception to this was the Army-wide and MOS-specific PRS, which exhibited very low interrater reliability coefficients. Revisions to the measures intended to improve their reliability are underway and will be presented in more detail in the next TOPS IOT&E evaluation report. Until improvements are implemented and reflected in the analysis data files, results concerning these scales should be interpreted with caution.

CHAPTER 5: VALIDITY RESULTS AND COMPOSITE FORMATION

Matthew C. Reeder, Matthew T. Allen, and Michael J. Ingerick (HumRRO)

In this chapter, we begin with a brief description of the current TOPS composites and how they were developed. This is followed by analyses examining the TAPAS' potential to enhance Soldier selection in two samples of interest. Next, we refer to analyses conducted to develop revised composites to replace the current composites. Due to sensitivity concerns, the results of these composite formation analyses are presented in separately-published appendices.

Background and Approach

As described in Chapter 3, when Army applicants take the TAPAS, their facet scores are averaged into two TOPS composites developed as part of ARI's EEEM research project (Knapp & Heffner, 2010). Based on the Campbell, Hanson, and Oppler (2001) job performance framework, the "can-do" composite consists of five TAPAS scales designed to predict Soldier performance on technical or job-specific criteria such as job knowledge. The "will-do" composite also consists of five TAPAS scales and is designed to predict less technical and job-specific dimensions of performance such as physical fitness, as well as retention-related criteria such as attrition and adjustment to the Army.

Previous TOPS IOT&E evaluations (Knapp & Heffner, 2011; Knapp, Heffner, et al., 2011) have found that (a) the descriptive statistics for some of the TAPAS scales that constitute the can-do and will-do composites changed from the research setting to the IOT&E setting (Allen et al., 2011) and (b) the validity coefficients for many of those same scales became non-significant in the IOT&E context (Caramagno et al., 2011). Meanwhile, other TAPAS scales not included in the original TOPS composites were more predictive of key criteria of interest than others that were included. These results suggest that the current TOPS composites should be revised based on the IOT&E data.

Evaluating the Predictive Potential of the TAPAS

Predictive Potential of the TAPAS in the Full Validation Sample

Approach

Our approach to analyzing the TAPAS' predictive validity in the TOPS sample was consistent with previous evaluations of the measure or similar experimental non-cognitive predictors (Ingerick et al., 2009; Knapp & Heffner, 2009, 2010; Trippe, Caramagno, et al., 2011). In brief, this approach involved testing a series of hierarchical regression models, regressing each criterion measure onto Soldiers' AFQT scores in the first step, followed by their TAPAS scores (i.e., the 15 facet scales) in the second step. The resulting increment in the multiple correlation (ΔR) when the TAPAS scores were added to the baseline regression models served as our index of incremental validity.

For the continuously scaled criteria, these models were estimated using Ordinary Least Squares (OLS) regression. Logistic regression was used for dichotomous criteria (e.g., 3- and 6-

month attrition). At each step in the model, we estimated point-biserial correlations (r_{pb}) in place of the traditional pseudo *R* estimates to index incremental validity because of conceptual and statistical issues associated with these estimates. The point-biserial correlations reflected the correlation between a Soldiers' predicted probability of engaging in a behavior based on the predictors in the logistic regression model and their actual behavior (e.g., attrition).

In addition to these incremental validity analyses, we examined the predictive validity of the TAPAS at the scale level using bivariate correlations. These results are reported in Appendix C.

Findings

The results of the incremental validity analyses can be found in Table 5.1. We report the results separately for Education Tier 1 and Tier 2 applicants, given current differences in the screening methods for the two samples (White et al., 2004). However, given the large number of predictors (i.e., 15 TAPAS scales plus AFQT) and the small sample sizes for the Education Tier 2 sample, the results are too unstable to be interpretable. While these results are included in Table 5.1 for the sake of completeness, the remainder of our interpretation will focus on Education Tier 1 TOPS Applicants.

Consistent with previous research, the TAPAS was generally more predictive of will-do performance and retention-related criteria than can-do performance-related criteria beyond AFQT (see Appendix C for the zero-order correlations). Across all criteria, the TAPAS was most predictive of Soldier physical fitness (Last APFT Score $\Delta R = .21$; Fitness and Bearing PRS $\Delta R = .12$), attitudes towards the Army (Adjustment to Army Life $\Delta R = .16$; Army Fit $\Delta R = .16$; Affective Commitment and Commitment PRS $\Delta R = .12$), and number of training restarts (ALQ Training Restarts $\Delta R = .17$). In spite of the very low interrater reliability coefficients (see Chapter 4), the TAPAS was a statistically significant predictor of all of the PRS. Though the magnitude of the coefficients was small for Tier 1 Soldiers, the TAPAS was also a statistically significant predictor of 3- and 6-month attrition. The magnitude of these effects was very similar to those found in the last evaluation cycle (Caramagno et al., 2011).

Table 5.1. Incremental Validity Estimates for the TAPAS Scales over the AFQT for Predicting Performance- and Retention-Related Criteria

	Tier 1				Tier 2			
Criterion	*n*	AFQT Only *R* (r_{pb})	AFQT + TAPAS *R* (r_{pb})	Δ*R* (Δr_{pb})	*n*	AFQT Only *R* (r_{pb})	AFQT + TAPAS *R* (r_{pb})	Δ*R* (Δr_{pb})
Can-Do Performance								
WTBD JKT	4,310	**.49**	.49	**.00**	108	**.31**	.47	.16
MOS-Specific JKT	3,634	**.38**	.39	**.01**	88	**.28**	.52	.25
MOS-Specific PRS	1,282	.04	.13	.10	31	.12	.93	**.81**
IMT Exam Grade	7,040	**.30**	.31	**.01**	158	**.27**	.40	.13
Graduated IMT without Restart (Academic)	15,114	(.00)	(.07)	(**.07**)	436	(.04)	(.15)	(.12)
Training Achievement	4,428	**.12**	.22	**.10**	111	**.19**	.48	.29
Training Restarts	4,438	**.05**	.22	**.17**	111	.03	.34	.31
Can-Do PRS	1,530	.03	.14	**.11**	36	.25	.83	**.58**
Will-Do Performance								
Effort and Discipline PRS	1,544	**.07**	.17	**.10**	36	.26	.77	.50
Fitness and Bearing PRS	1,536	**.09**	.21	**.12**	36	.18	.76	.59
Work with Other PRS	1,533	.04	.15	**.12**	36	.23	.72	.49
Last APFT Score	4,387	**.10**	.31	**.21**	108	.13	.54	.41
Disciplinary Incidents	2,904	.03	.11	**.09**	66	.08	.54	.45
Overall Performance	1,534	**.07**	.17	**.10**	35	.23	.78	.54
Retention								
Affective Commitment	4,438	**.09**	.21	**.12**	111	.08	.39	.32
Attrition Cognitions	4,438	.02	.16	**.14**	111	.05	.43	.38
Adjust to Army Life	4,438	**.07**	.23	**.16**	111	.05	.44	.39
Commit and Adjust PRS	1,541	.03	.15	**.12**	36	.19	.77	.59
Army Fit	4,438	**.04**	.20	**.16**	111	.06	.42	.36
MOS Fit	4,438	**.04**	.16	**.12**	111	.05	.28	.23
3-Month Attrition[a]	23,101	(**.04**)	(.08)	(**.04**)	270	(.03)	(.36)	(**.34**)
6-Month Attrition[a]	17,541	(**.06**)	(.11)	(**.05**)	219	(.01)	(.27)	(.27)
9-Month Attrition[a]	10,249	(**.07**)	(.11)	(.04)	155	(.02)	(.42)	(.40)

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. PRS = Performance Rating Scales. AFQT Only = Correlation between the AFQT and the criterion of interest. AFQT + TAPAS = Multiple correlation (*R*) between the AFQT and the selected predictor measure with the criterion of interest. Δ*R* = Increment in *R* over the AFQT from adding the selected predictor measure to the regression model ([AFQT + TAPAS]—AFQT Only). Estimates in parentheses are *point-biserial correlations* (r_{pb}) that reflect the observed *point-biserial correlation* (r_{pb}) between Soldiers' predicted probability of an event (e.g., attrition, graduating IMT without a restart) and their actual behavior. Large, positive r_{pb} values mean that the TOPS composite or scale performed well in predicting the target outcome. Results are limited to non-prior service, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

These results suggest that the TAPAS has promise for predicting multiple criteria of interest in an applicant setting. As noted above, however, the target sample for the new TOPS composites is restricted to Tier 1, AFQT Category IIIB and IV applicants within the TOPS Validation Sample. It is this subset of the sample that is the subject of the remainder of the chapter.

Predictive Potential of TAPAS Scales in the Restricted (AFQT Category IIIB and IV) Sample

Approach

The approach taken to examining the predictive potential of TAPAS in the restricted sample mirrors that taken in the broader Validation Sample, with three exceptions. First, given the smaller sample sizes and the number of predictors included in the regression models, sample-specific error is a greater concern in the restricted sample than in the full TOPS Validation sample. To account for this, we corrected the estimated multiple *R*'s for population cross-validity using Burket's formula, described in Schmitt and Ployhart (1999; see Formula 8; denoted by ρ_c). Second, to apply the Burket formula to dichotomous outcomes, Cox and Snell's pseudo estimate of *R* was used in place of the r_{pb} procedure used in the previous section. Third, given that the composite analysis sample is already restricted to Tier 1, AFQT Category IIIB and IV applicants, AFQT was not included in the regression models. Table 5.2 shows multiple correlations and cross-validity estimates for each criterion regressed on the 15 TAPAS scales.

Findings

Of the can-do criteria, the best-predicted outcomes included MOS-Specific PRS (*R* = .28), the Can-Do PRS (*R* = .26), Training Achievements (*R* = .19), and Training Restarts (*R* = .17). The estimated cross-validities for these estimates were comparable to the raw estimates found in the TOPS Applicant Sample (see Table 5.1). Of the will-do criteria, the most well-predicted outcomes included Effort and Discipline PRS (*R* = .29) and Last APFT Score (*R* = .27). The most well-predicted retention-related criteria included Affective Commitment and Army Fit (*R* = .23), Attrition Cognitions (*R* = .20), Adjustment to Army Life (*R* = .21), and MOS Fit (*R* = .18). As with the can-do criteria, the magnitude of the estimated cross-validities for these criteria was comparable to or higher than what was found in the Validation Sample. Generally, the cross-validity estimates were all positive, though a few exceptions did occur for criteria that were less well-predicted by the TAPAS (notably, WTBD JKT and Disciplinary Incidents).

These results suggest that the pattern and magnitude of prediction in this subsample is comparable to what was found in the full Applicant Sample. The next step in the analysis was to develop and evaluate new TOPS composites designed specifically for predicting can-do and will-do/retention outcomes of relevance to the Army.

Constructing and Evaluating Revised TOPS Composites

HumRRO and DCG each conducted follow-up analyses to provide independent approaches to (a) develop new predictor composites and (b) evaluate the new composites against the old composites in terms of predictive utility. Each had different parameters from which to work (e.g., the number of composites to be generated) representative of different implementation scenarios. Although they each used different assumptions, both the HumRRO and DCG-developed composites represented a substantial improvement in terms of predictive efficacy over the original can-do and will-do composites, particularly for the restricted (Tier 1, AFQT Category IIIB and IV) sample. Descriptions of the approach and results of these analyses are

provided in Appendices D and E. Those interested in obtaining a copy of these appendices should contact the editors for further information.

Table 5.2. Validity Estimates for the TAPAS Scales for Predicting Performance- and Retention-Related Criteria for Education Tier 1, AFQT Category IIIB and IV Soldiers

Predictor/Scale	*n*	*R*	ρ_c
Can-Do Performance			
WTBD JKT	1,203	.11	-.01
MOS-Specific JKT	935	.15	.05
MOS-Specific PRS	334	**.28**	.13
IMT Exam Grade	2,449	**.10**	.04
Graduated IMT without Restart (Academic)	5,136	**.07**	.03
Training Achievement	1,241	**.19**	.12
Training Restarts	1,242	**.17**	.10
Can-Do PRS	366	**.26**	.11
Will-Do Performance			
Effort and Discipline PRS	367	**.29**	.15
Fitness and Bearing PRS	365	.26	.10
Work with Other PRS	364	.26	.10
Last APFT Score	1,228	**.27**	.22
Disciplinary Incidents	833	.12	-.04
Retention			
Affective Commitment	1,242	**.23**	.18
Attrition Cognitions	1,242	**.20**	.14
Adjust to Army Life	1,242	**.21**	.15
Commit and Adjust PRS	367	.25	.09
Army Fit	1,242	**.23**	.18
MOS Fit	1,242	**.18**	.12
3-Month Attrition[a]	7,754	**.07**	.04
6-Month Attrition[a]	5,895	**.08**	.06
9-Month Attrition[a]	3,432	**.09**	.03
Overall Performance	366	.22	.03

Note. Estimates in bold are statistically significant at the $p < .05$ level.

[a]Model *R* estimates for the dichotomous criteria were computed as $\sqrt{R^2_{CS}}$, where R^2_{CS} is the Cox and Snell pseudo-R^2 model estimate.

Summary

We examined the validity of TAPAS for incrementally predicting various outcomes of interest over the AFQT in the Education Tier 1 Validation sample. Consistent with previous reports in this stream of research (Knapp et al., 2010; Knapp & Heffner, 2010; 2011), the TAPAS was most predictive of will-do performance criteria, such as physical fitness and effort, and retention-related criteria, such as commitment and adjustment. In contrast, the TAPAS was generally less predictive of can-do performance criteria, such as job knowledge. When conducting similar analyses in a more restricted sample (limited to applicants in AFQT Categories IIIB and IV), the TAPAS remained predictive of multiple outcomes of interest, even when controlling for the number of predictors in the regression model and sample size. Overall, these results suggest that the TAPAS is a promising instrument for enhancing the Army's procedures for selecting Soldiers. The composites developed independently by HumRRO and DCG target these Tier 1, AFQT Category IIIB and IV Soldiers, and reflect the Army's current recruiting, selection, and accessioning environment. Should these conditions change, additional flexibility may be afforded to the composite development procedures in the future that better maximizes the operational use of the TAPAS in a selection environment.

CHAPTER 6: EVALUATION OF TAPAS POTENTIAL FOR CLASSIFICATION PURPOSES

Matthew Trippe, Ted Diaz, and Michael Ingerick (HumRRO)

Introduction

Similar to previous research (Ingerick et al., 2009; Knapp, Heffner, et al., 2011), we evaluated the experimental predictor measures' classification potential using (a) Horst's (1954, 1955) index of differential validity (H_d) and (b) Brogden's expected criterion scores of optimally assigned individuals (De Corte, 2000). Conceptually, H_d provides an index of the predictor measure(s)' ability to differentiate among the predicted criterion scores for a sample of jobs. The greater the H_d value, the larger the cross-job differences in the predicted criterion scores. Analytically, H_d represents the average standardized mean difference between all possible pairs of predicted criterion scores for a sample of jobs. Conversely, Brogden's expected criterion scores reflect the predicted criterion scores for Soldiers optimally assigned to a sample of jobs using the predictor measures. A common way to summarize predicted criterion scores is simply with the mean predicted criterion score (MPCS). The greater the MPCS, the higher Soldiers are predicted to perform or be satisfied, on average, when classified into a sample of jobs using the selected predictor measures. However, expected criterion scores are traditionally expressed in a standardized metric with a known distribution that is common across MOS. We report results in the metric of the criterion being analyzed to make interpretation of outcomes less abstract. Thus, predicted criterion scores presented here are best interpreted in terms of distributional properties (e.g., means and percentiles). Interpreting the MPCS in the context of additional distributional properties provides a more complete and accurate picture of the classification context.

Although the two classification indices are related (i.e., larger H_d values tend to be associated with higher MPCS values), each captures unique information about the classification potential of the predictor measures. Whereas H_d provides information on cross-job differences (or variability) in Soldiers' predicted criterion scores resulting from the use of the predictor measures to classify Soldiers into a sample of jobs, the MPCS supplies information on the average level at which Soldiers are predicted to score on the targeted criterion (e.g., performance, retention). H_d can be viewed as somewhat of a descriptive or diagnostic indicator of classification potential that does not include all of the factors modeled in Brogden's expected criterion scores. The latter index is a more comprehensive index that accounts for a number of additional factors, including the percentage of Soldiers allocated to each MOS and the optimal assignment of each Soldier to an MOS with respect to the criterion being analyzed. Brogden's index, like H_d, considers the degree of differential validity among predictor composites when attempting to optimize classification.

Approach to Estimating the Classification Potential

Comparable to the incremental predictive validity analyses, we estimated the increment in H_d and MPCS resulting from using the TAPAS over existing ASVAB subtests[14] to enhance new Soldier classification. Consistent with the Army's personnel management objectives, we investigated the measures' potential for enhancing both performance and retention-related criteria.

[14] General Science (GS), Arithmetic Reasoning (AR), Math Knowledge (MK), Electronics Information (EI), Auto Shop (AS), Mechanical Comprehension (MC), Verbal composite (VE) of Word Knowledge (WK) and Paragraph Comprehension (PC). Assembling Objects (AO) was not included because (a) it is not currently part of any existing Aptitude Area composites and (b) missing data are prevalent in this subtest.

Previous research (Ingerick et al., 2009; Knapp, Heffner, et al., 2011) examined continuous criterion scores, which were assumed to be linearly related to the ASVAB and experimental predictor measures. In the current work, we expanded Brogden's classification framework to handle dichotomous criterion scores. In general, the steps in the expanded approach closely follow those for the continuous criterion scores. The main difference between the traditional and expanded approach is in how the underlying classification composites used for optimally assigning individuals to jobs are obtained. In the traditional approach, with continuous criterion scores assumed to be linearly related to the predictors, the classification composite is obtained as the linear function of the predictors using multiple regression methods (e.g., least squares estimation or conditional normal regression). In the expanded approach, we assume a logistic probability model to relate the dichotomous criterion response to the predictors. We then transformed the predicted probabilities using the logit function to obtain a linear composite of predictors needed in DeCorte's (2000) multivariate normal formulation of Brogden's classification framework.

Our analysis approach thus consisted of the following general steps.

1. Estimate the linear predictor composite for each MOS.
2. Estimate the observed (uncorrected) predictor-linear composite covariance matrix for each MOS.
3. Correct the predictor-linear composite covariances from Step 2 for multivariate range restriction on the ASVAB and TAPAS using the entire "accession sample" as the reference population (Lawley, 1943). [15]
4. Using the corrected predictor-linear composite covariance from Step 3, compute the multiple correlation of the linear composites.
5. Correct the multiple correlations of linear composites for cross-validity (Burket, 1964).
6. Using the corrected covariance matrices from Step 5, compute two indices of classification potential: (a) (H_d) and (b) Brogden's expected criterion scores of optimally assigned individuals (DeCorte, 2000).

Several factors should be kept in mind when interpreting these results. First, our analyses did not model important organizational factors and other operational constraints that contribute to the Soldier-job matching process under the Army's existing classification system (e.g., demand for certain MOS, availability of training seats at the time of accession). Classification models include all of the ASVAB (except AO) and TAPAS subtest predictors. Although this allows us to address the issue of classification potential, it also allows for predictors to be used in an optimal fashion that does not reflect practical operational usage. Including nearly all subtests allows the ASVAB to account for more variance than it would operationally. The TAPAS scales are also used in this optimal manner to provide a balanced evaluation. Accordingly, the estimates reported reflect the *potential* of the predictor measures to enhance new Soldier classification and not the *actual* expected gains in classification if the measures were used operationally. Second, the results reported could differ if a different sample of MOS or set of criterion measures were

[15] The "accession sample" includes Soldiers from the Applicant Sample who signed an enlistment contract.

examined. Accordingly, we focused our analyses on the MOS targeted for the TOPS project. Focusing on these MOS (or subsets of them as sample sizes permit) ensures that analysis results remain generally comparable. Third, there are no standards or conventions for interpreting the magnitude of or gain in H_d relative to some baseline. There is some evidence that increments in MPCS as low as .10 carry significant and practical operational gains (Nord & Schmitz, 1991). Past research examining the Project A experimental predictor measures found increments in MPCS ranging from .05 to .10 when the selected experimental predictors were combined with the ASVAB to maximize a performance-based criterion (Rosse, Campbell, & Peterson, 2001; Scholarios, Johnson, & Zeidner, 1994). Nevertheless, those MPCS values are reported in a standardized metric and results presented below are in the metric of the criterion analyzed. These results are best interpreted in terms of the relative improvements in distributions of criterion scores. That is, the less overlap there is between the distributions of predicted criterion scores, the greater the classification improvement.

Results

Table 6.1 provides a summary of the overall classification potential indices by criterion measure. Criterion measures were selected based on both expectations of cross-MOS differences as well as availability of data in MOS analyzed. Attrition and IMT restarts are outcome measures for which we have relatively large samples and where we might expect to see cross-MOS differences resulting from variation in training demands. MOS-specific JKTs and ALQ scales were chosen based on the expectation of cross-MOS differences resulting from variation in training demands and experiences. Table 6.1 contains H_d and MPCS values for a predictive model that includes the ASVAB subtests and a model that includes the ASVAB subtests as well as the TAPAS scales. MPCS values presented in Table 6.1 are overall means computed across MOS and weighed by the MOS allocation percentages in each model (see notes in Tables 6.2 through 6.6 for these percentages). H_d values indicate that there is relatively more variability in predicting the attrition and IMT restart criterion variables across MOS when the TAPAS is added to the model. Conversely, there is relatively little variability in predicting the MOS specific JKT, Army Life Adjustment or MOS Fit criterion variables across MOS in either model. Overall, MPCS values demonstrate modest improvements when the TAPAS scales are added to a model predicting the attrition and IMT restart criterion variables. Little or no increment in MPCS is observed when the TAPAS scales are added to the model predicting the MOS-specific JKT, Army Life Adjustment or MOS Fit criterion variables. However, fewer MOS and smaller sample sizes are currently available for those criterion variables.

Tables 6.2 through 6.6 contain MPCS values as well as the 5^{th}, 50^{th}, and 95^{th} percentile of predicted criterion scores by MOS and averaged across MOS. Percentiles are reported to provide a sense of lower, middle, and upper portions of the distribution of predicted criterion values, which are reported in the metric of the criterion variable being analyzed. As in the summary table, values are presented for a predictive model that includes the ASVAB subtests alone and a model with both the ASVAB subtests and the TAPAS scales. Bolded values within each table represent instances where the model including the TAPAS is a significant improvement over the model that includes the ASVAB subtests alone. “Significant” is defined here as when the distribution of predicted scores do not overlap across the two models.

Table 6.1. Summary of Overall Classification Potential Indices of the TAPAS Relative to the ASVAB by Criterion Measure

Criterion		H_d		*MPCS*	
	Number of MOS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS
3-Month Attrition	8	.038	.149	5.9 %	4.8 %
6-Month Attrition	8	.029	.142	9.7 %	8.0 %
IMT Restart	8	.097	.246	13.7 %	12.6 %
MOS-Specific JKT	5	.027	.039	65.1 %	65.4 %
Army Life Adjustment	6	.006	.030	4.083	4.094
MOS Fit	6	.033	.031	3.791	3.775

Note. Number of MOS refers to how many MOS were included in the classification analysis. H_d represents the average standardized mean difference between all possible pairs of predicted criterion scores for a sample of jobs. MPCS represents the mean predicted criterion scores for Soldiers optimally assigned to a sample of jobs using the predictor measures. MPCS values are reported in the metric of the criterion measure, so lower values are better for attrition and IMT restart, while higher values are better for the JKT, Army Life Adjustment, and MOS Fit.

Table 6.2. Classification Potential of the TAPAS Relative to the ASVAB for Minimizing 3-Month Attrition

	Predicted Criterion: Percent 3-Month Attrition							
	Mean		*5th Percentile*		*50th Percentile*		*95th Percentile*	
	ASVAB Only	ASVAB+ TAPAS	ASVAB Only	ASVAB+ TAPAS	ASVAB Only	ASVAB+ TAPAS	ASVAB Only	ASVAB+ TAPAS
Overall	5.9	4.8	3.4	0.6	5.8	4.9	7.9	8.6
11B	7.6	7.7	7.1	6.4	7.6	7.7	8.0	9.0
19K	1.1	**0.1**	0.8	0.0	1.1	0.1	1.4	0.1
25U	4.6	**0.9**	3.8	0.4	4.6	0.9	5.4	1.7
31B	4.1	**1.8**	3.5	1.1	4.1	1.8	4.7	2.8
42A	3.7	**0.9**	3.0	0.4	3.6	0.8	4.5	1.5
68W	5.6	4.6	4.9	3.4	5.7	4.6	6.4	5.9
88M	4.8	3.2	4.3	2.4	4.8	3.2	5.3	4.2
91B	4.7	**1.3**	3.9	0.6	4.7	1.3	5.5	2.1

Note. Values in the table represent summary statistics of predicted rates of attrition when the ASVAB or the ASVAB + TAPAS are used to classify individuals into the MOS listed. Lower values indicate lower predicted rates of attrition. Percentiles are reported to provide a sense of lower, middle, and upper portions of the distribution of predicted criterion values. Allocation percentages (11B = 45%, 19K=3%, 25U=4%, 31B=9%, 42A =4%, 68W=14%, 88M = 10%, 91B =10%) are based on the number of Soldiers in each MOS in the TOPS "accession sample." Classification estimates are derived from the "accession sample" Soldiers who have non-missing predictor and criterion data (11B n = 5,671, 19K n = 418, 25U n = 330, 31B n = 477, 42A n = 220, 68W n = 1,344, 88M n = 819, 91B n = 796). Bolded values are those where the distributions of predicted criterion scores based on the ASVAB and ASVAB + TAPAS do not overlap and thus represent significant improvement in predicted outcomes.

Table 6.2 shows that the mean predicted 3-month attrition rate across all eight MOS included in the analysis is 5.9% when the ASVAB subtests are used to classify individuals into MOS. When the TAPAS is added to the ASVAB subtests in the classification model, this overall attrition rate falls to 4.8%. Although an overall reduction in 3-month attrition of just 1% is certainly modest, evaluation of the distribution of MOS 3-month attrition rates reveals more substantive improvements. For example, the mean predicted 3-month attrition rate for 31B is 4.1% when the ASVAB alone is used to classify, but this rate is reduced to 1.8% when the TAPAS is added to the classification model. That is, the mean predicted 3-month attrition rate is roughly reduced by half for this MOS.

Figure 6.1 provides a visual representation of the distributional properties expressed in Table 6.2. The boxplots display the mean, median, and interquartile range as well as the 5th and 95th percentiles of predicted attrition percentages by MOS. The dark grey boxes represent the distribution of predicted attrition percentages when the ASVAB subtests serve as predictors. The light grey boxes represent the distribution of predicted attrition percentages when the ASVAB and TAPAS serve as predictors. The pattern of results reveals that the distribution of 3-month attrition rates for most MOS is reduced by adding the TAPAS to the classification model that includes the ASVAB. That is, the light grey boxplots associated with the ASVAB + TAPAS model generally demonstrate that the distribution of 3-month attrition is lower on the Y-axis than the dark grey boxplots representing the ASVAB subtests alone.

Table 6.2 and Figure 6.1 also reveal that the attrition rate for 11B remains virtually unchanged when the TAPAS is added to a classification model containing the ASVAB. There are a few possible explanations for this finding. First, the classification model must allocate nearly half (45%) of individuals to this MOS. Because 11B represents such a large proportion with respect to the other MOS, the model cannot be as selective on who is classified into this MOS. Similarly, 11B attrition is harder to predict than other MOS in this model. The overall amount of variance explained in attrition by the ASVAB and the TAPAS is lower in 11B than for the other MOS.[16] Thus, when the model tries to optimize (reduce) attrition in this group of MOS, it is difficult to achieve gains for 11B. Stated more generally, MOS with relatively small allocation percentages and relatively high amounts of variance accounted for in predicting outcomes can gain the most in the classification model. In the analysis presented in Table 6.2, 11B has both a high allocation percentage and relatively low amount of variance explained in attrition. Moreover, we applied a correction for cross validity (Burket, 1964) to the multiple correlations that are part of the basis for the classification estimates, which will penalize the less parsimonious model including both the ASVAB and TAPAS scales.

Table 6.3 presents the distributions of predicted 6-month attrition rate for the eight MOS included in the analysis. As in the 3-month attrition analyses, when the overall incremental reduction in attrition achieved by adding the TAPAS scales to the model is averaged across all MOS, the improvement is rather small. Nevertheless, there are a number of MOS for which the predicted 6-month attrition rate is significantly reduced. For example, the average predicted 6-month attrition rate for 19K is 8.1% when the ASVAB subtests alone are used to classify individuals, but that average predicted rate is reduced to 2.7% when the TAPAS is added to the model. Again, the MOS that see the greatest improvements are those that are both well predicted and have relatively smaller allocation ratios. Conversely, the distribution of predicted 6-month attrition rate for the 11B sample actually rises slightly when TAPAS is added to the model that contains ASVAB. In this classification scenario where every individual must be assigned to an MOS according to the allocation percentages, the gains observed in the MOS whose rates are substantially reduced come at the expense of the MOS with larger allocation percentages and relatively weaker predictor-criterion relationships.

[16] Multiple correlations are part of a larger matrix that serves as input to the classification model and are not reported in Tables 6.2 through 6.6.

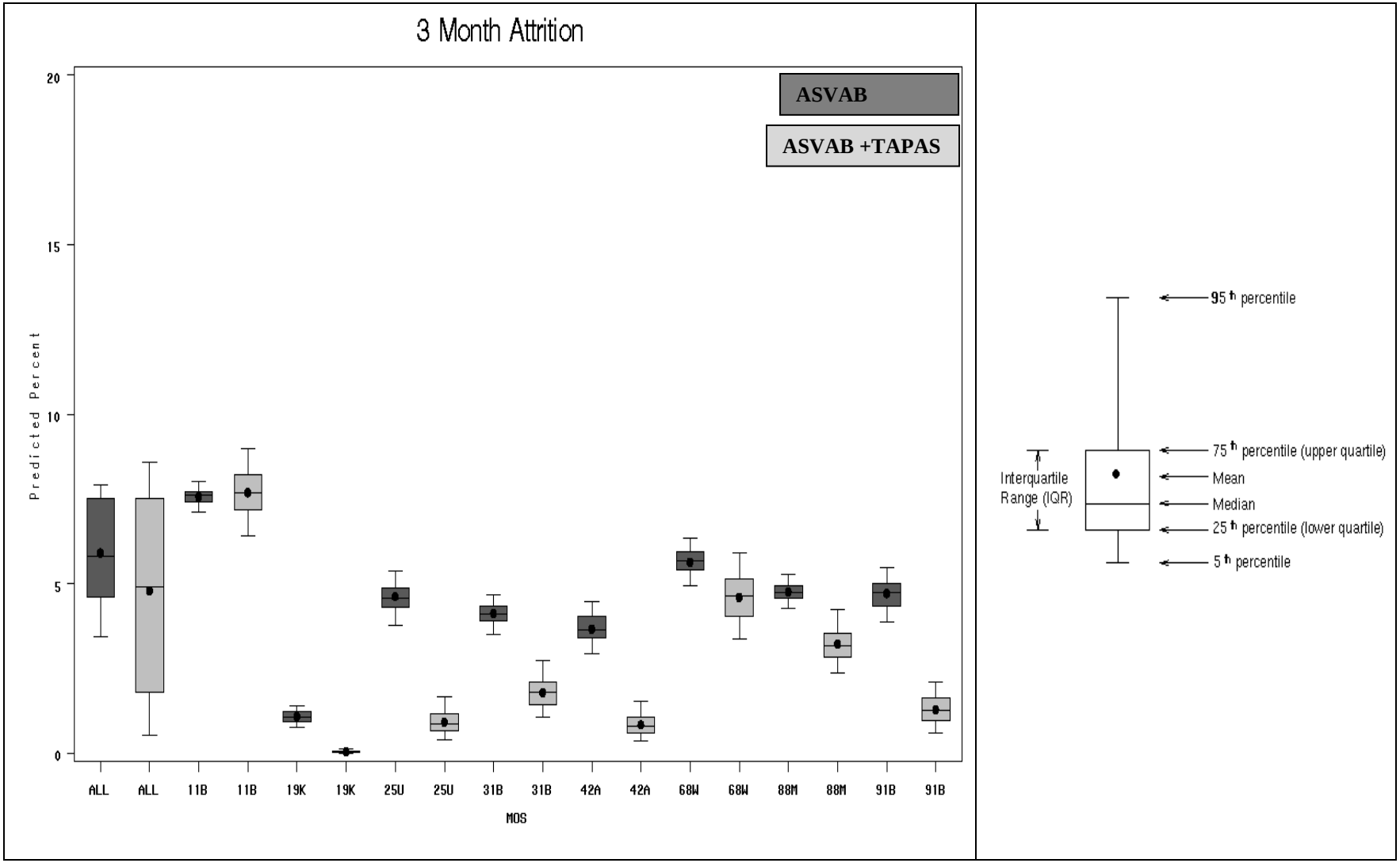

Figure 6.1. Boxplot Presentation of the Classification Potential of the TAPAS Relative to the ASVAB for Minimizing 3-Month Attrition.

Table 6.3. Classification Potential of the TAPAS Relative to the ASVAB for Minimizing 6-Month Attrition

	Predicted Criterion: Percent 6-Month Attrition							
	Mean		*5th Percentile*		*Median*		*95th Percentile*	
	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS
Overall	9.7	8.0	6.1	0.7	9.5	7.9	12.0	13.5
11B	11.5	12.0	11.0	10.1	11.6	12.1	12.0	14.1
19K	8.1	**2.7**	7.5	1.4	8.2	2.6	8.8	4.2
25U	8.2	**1.1**	7.6	0.5	8.1	1.0	8.7	1.9
31B	9.4	6.8	8.4	4.8	9.3	6.6	10.3	9.0
42A	5.0	**0.1**	4.2	0.0	5.0	0.1	5.6	0.3
68W	8.4	7.1	7.7	5.5	8.5	7.1	9.0	8.8
88M	9.1	**5.6**	8.5	4.0	9.2	5.6	9.7	7.3
91B	7.1	**3.0**	6.0	1.6	7.1	2.9	7.9	4.5

Note. Values in the table represent summary statistics of predicted rates of attrition when the ASVAB or the ASVAB + TAPAS are used to classify individuals into the MOS listed. Lower values indicate lower rates of attrition. Percentiles are reported to provide a sense of lower, middle, and upper portions of the distribution of predicted criterion values. Allocation percentages (11B = 45%, 19K=3%, 25U=4%, 31B=9%, 42A =4%, 68W=14%, 88M = 10%, 91B =10%) are based on the number of Soldiers in each MOS in the TOPS "accession sample". Classification estimates are derived from the "accession sample" Soldiers who have non-missing predictor and criterion data (11B n = 4,601, 19K n = 280, 25U n = 208, 31B n = 258, 42A n = 165, 68W n = 960, 88M n = 653, 91B n = 625). Bolded values are those where the distributions of predicted criterion scores based on the ASVAB and ASVAB + TAPAS do not overlap and thus represent significant improvement in predicted outcomes.

Table 6.4 presents the distributions of predicted percentage of IMT restarts in the eight target MOS. Although a number of the MOS level results suggest appreciable improvements, it must be noted that the combination of smaller sample sizes and the relative infrequency of restart events in these MOS results in some unstable results. Predicted restart rates achieve values in 19K, 25U, and 42A that are likely to be unrealistic. For example, the predicted restart rate for 19K in the ASVAB-only classification model is 9.9, but is reduced to effectively zero when TAPAS is added to the model. Predicted restart rates achieve values of zero or nearly zero in 25U and 42A as well. It may be tempting to believe that the classification model was overwhelmingly successful in these instances, but the reality is that many of these results are likely artifacts of insufficient data. We expect this analysis to stabilize as more data accumulate. These preliminary results suggest promise in the capacity for the TAPAS to achieve incremental classification gains beyond the ASVAB subtests with respect to this criterion.

Table 6.5 contains the distributions of predicted percentage correct on the MOS-specific JKTs for the five MOS that have at least 100 Soldiers with criterion data. The results suggest that there is virtually no change in the predicted distribution of JKT scores when the TAPAS scales are added to a model including the ASVAB subtests. The H_d associated with the MOS-specific JKT found in Table 6.1 provides context here, in that it indicates there is not much variability in the predicted JKT scores across MOS. Thus, it is difficult for the classification model to optimally assign Soldiers to MOS when there is not much difference in the predicted outcomes across occupations. This may be due in part to the reduced number of MOS in this analysis with available criterion data. More directly related to the lack of incremental gain in classification potential of the TAPAS over the ASVAB in this model is lack of incremental prediction. That is, the TAPAS does not explain incremental variance beyond the ASVAB in the prediction of MOS-specific JKT

scores. This is not altogether surprising given the highly cognitive nature of both the ASVAB and JKTs. Factors working against the TAPAS scales in this classification model are the non-cognitive nature of the predictors and the shrinkage correction. The correction for cross-validation will offset any modest incremental prediction gains achieved by the TAPAS in this model because of the lack of parsimony associated with the relatively large number of predictors.

Table 6.4. Classification Potential of the TAPAS Relative to the ASVAB for Minimizing IMT Restart

	Predicted Criterion: Percent with at Least One IMT Restart							
	Mean		*5th Percentile*		*Median*		*95th Percentile*	
	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS
Overall	13.7	12.6	1.5	0.0	15.8	15.1	20.5	20.5
11B	15.8	16.1	15.2	14.0	15.8	16.2	16.3	18.3
19K[a]	9.9	**0.1**	8.5	0.0	9.9	0.1	11.2	0.3
25U[a]	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0
31B	8.9	8.6	7.7	6.7	8.8	8.5	10.1	10.9
42A[a]	1.8	**0.1**	1.3	0.0	1.8	0.1	2.2	0.3
68W	19.9	19.7	17.4	15.5	20.1	19.9	22.0	23.8
88M	16.5	13.9	16.0	11.8	16.6	14.0	16.9	16.1
91B	9.6	**3.9**	8.8	2.2	9.6	3.9	10.2	6.1

Note. Values in the table represent summary statistics of predicted rates of restart when the ASVAB or the ASVAB + TAPAS are used to classify individuals into the MOS listed. Lower values indicate lower rates of restart. Percentiles are reported to provide a sense of lower, middle, and upper portions of the distribution of predicted criterion values. Allocation percentages (11B = 45%, 19K=3%, 25U=4%, 31B=9%, 42A =4%, 68W=14%, 88M = 10%, 91B =10%) are based on the number of Soldiers in each MOS in the TOPS "accession sample." Classification estimates are derived from the "accession sample" Soldiers who have non-missing predictor and criterion data (11B n = 4,170, 19K n = 164, 25U n = 117, 31B n = 440, 42A n = 315, 68W n = 565, 88M n = 1030, 91B n = 554). Bolded values are those where the distributions of predicted criterion scores based on the ASVAB and ASVAB + TAPAS do not overlap and thus represent significant improvement in predicted outcomes.

[a] Sample sizes for 19K, 25U, and 42A in combination with the relative infrequency of failure events in these MOS results in some unstable/unrealistic results in this analysis. We expect this analysis to stabilize as more data accumulate.

Table 6.5. Classification Potential of the TAPAS Relative to the ASVAB for Maximizing MOS Specific JKT scores

	Predicted Criterion: MOS Specific JKT Percent Correct							
	Mean		*5th Percentile*		*Median*		*95th Percentile*	
	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS
Overall	65.1	65.4	56.5	56.5	62.6	63.2	77.0	76.3
11B	61.5	61.9	58.9	58.9	61.5	62.0	63.8	65.1
31B	77.1	76.2	74.4	72.5	76.7	76.0	80.7	80.4
68W	70.3	70.8	67.7	68.3	70.4	70.8	73.0	73.2
88M	71.6	71.7	67.8	66.9	71.5	71.6	76.0	76.6
91B	57.1	57.4	51.0	51.2	56.9	57.5	63.2	63.9

Note. Values in the table represent summary statistics of predicted proportion correct on the MOS specific JKT when the ASVAB or the ASVAB & TAPAS are used to classify individuals into the MOS listed. Higher values indicate better performance. Percentiles are reported to provide a sense of lower, middle, and upper portions of the distribution of predicted criterion values. Allocation percentages (11B = 51%, 31B=10%, 68W=16%, 88M = 12%, 91B =12%) are based on the number of Soldiers in each MOS in the TOPS "accession sample." Classification estimates are derived from the "accession sample" Soldiers who have non-missing predictor and criterion data (11B n = 1,637, 31B n = 642, 68W n = 799, 88M n = 460, 91B n = 139).

Table 6.6 presents the distributions of predicted scores on the Army Life Adjustment and MOS fit scales in the ALQ. Similar to what is seen in the MOS JKT results, the predicted distributions do not differ much between the two models. That is, the TAPAS scales do not demonstrate incremental classification gains over the ASVAB with respect to the ALQ criterion variables analyzes. There are a number of factors working against the classification model in these analyses. First, the H_d values found in Table 6.1 indicate that there is not much variability across MOS in the predicted outcomes. This prohibits the classification model from finding ways of optimally sorting Soldiers into MOS for which they would best adjust or fit. Second, the baseline values of these two criterion variables tend to be high. That is, most Soldiers in most MOS tend to endorse relatively high levels of adjustment and fit. It is therefore difficult to achieve gains that appear appreciable in criteria that are relatively high to begin with. Finally, sample sizes for 42A and 91B are relatively small in this analysis and combine with the correction for cross-validation to penalize the model including the TAPAS for lack of parsimony.

Table 6.6. Classification Potential of the TAPAS Relative to the ASVAB for Maximizing Perceptions of Army Life Adjustment and MOS Fit

	Predicted Criterion: ALQ Army Life Adjustment Mean Response							
	Mean		*5th Percentile*		*Median*		*95th Percentile*	
	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS	ASVAB	ASVAB+ TAPAS
Overall	4.083	4.094	3.964	3.959	4.109	4.096	4.154	4.207
11B	4.119	4.102	4.104	4.048	4.120	4.095	4.135	4.161
31B	4.029	4.017	4.024	3.977	4.029	4.013	4.033	4.059
42A	4.085	4.233	4.083	4.171	4.083	4.232	4.093	4.309
68W	3.969	3.974	3.950	3.917	3.964	3.969	3.998	4.043
88M	4.155	4.163	4.154	4.129	4.154	4.164	4.158	4.198
91B	4.051	4.157	4.051	4.088	4.051	4.153	4.051	4.234
	Predicted Criterion: ALQ MOS Fit Mean Response							
Overall	3.791	3.775	3.211	3.201	3.909	3.877	4.001	4.005
11B	3.910	3.886	3.905	3.849	3.909	3.886	3.916	3.931
31B	3.872	3.843	3.859	3.816	3.872	3.845	3.886	3.873
42A	4.002	4.046	3.916	3.923	3.984	4.044	4.142	4.224
68W	3.997	3.986	3.990	3.958	3.995	3.986	4.006	4.014
88M	3.221	3.202	3.190	3.165	3.222	3.201	3.264	3.248
91B	3.405	3.396	3.364	3.326	3.395	3.390	3.474	3.476

Note. Values in the table represent summary statistics of predicted rates of adjustment/fit when the ASVAB or the ASVAB & TAPAS are used to classify individuals into the MOS listed. Higher values indicate better adjustment/fit. Percentiles are reported to provide a sense of lower, middle, and upper portions of the distribution of predicted criterion values. Allocation percentages (11B = 49%, 31B=9%, 42A=5% 68W=15%, 88M = 11%, 91B =11%) are based on the number of Soldiers in each MOS in the TOPS "accession sample." Classification estimates are derived from the "accession sample" Soldiers who have non-missing predictor and criterion data (11B n = 2,011, 31B n = 723, 42A n = 99, 68W n = 882, 88M n = 570, 91B n = 205).

Summary

The classification results presented in this chapter demonstrate that the TAPAS can provide incremental improvements beyond the ASVAB subtests for optimally assigning Soldiers to MOS. These incremental gains were observed in the dichotomous outcome variables (attrition and IMT restart) for which larger sample sizes in a greater number of MOS are available. In these cases, the reduction in predicted overall attrition or IMT restart is modest, but some MOS level results suggest a significant improvement. This is to some extent a result of the classification scenario modeled here, in which every Soldier must be assigned to an MOS according to the allocation percentages. An interesting alternative scenario for future research may be to introduce selection into the classification model. This allows for a specified percentage of Soldiers to remain unclassified or essentially rejected from assignment to any MOS in the analysis. That is not to say that such Soldiers would be rejected from the Army, but rather the implication is that they should be classified into an MOS other than those being considered for the particular analysis. Although this type of analysis is more of a theoretical exercise, it may free up some of the constraints that prevented the predictor composites from demonstrating their full potential.

Results from the JKT and ALQ criterion variables currently do not suggest that the TAPAS provides incremental improvements in classification beyond the ASVAB. Nevertheless, these analyses currently suffer from a number of limitations related to the availability of criterion data within and across the MOS. We expect these analyses to become more informative as these criterion data continue to accumulate and we obtain JKT and ALQ data for additional MOS.

CHAPTER 7: SUMMARY AND A LOOK AHEAD

Deirdre J. Knapp (HumRRO), Tonia S. Heffner, and Leonard A. White (ARI)

Summary of the TOPS IOT&E Method

The Army is conducting an IOT&E of the TOPS. The TOPS assessments, including the TAPAS, the ICTL test, and starting in CY2012, the WPA, are being administered to non-prior service applicants testing at MEPS locations.

To evaluate the TAPAS, ICTL, and WPA, the Army is collecting training criterion data on Soldiers in selected MOS as they complete their IMT. The criterion measures include JKTs, an attitudinal person-environment fit assessment (the ALQ), and PRS completed by the Soldiers' cadre members. Course grades and completion rates are obtained from administrative records for all Soldiers, regardless of MOS. The plan is to construct analysis datasets and conduct validation analyses at 6-month intervals throughout the IOT&E period.

At least two waves of in-unit job performance data collection are also planned at approximately 18 month intervals, each attempting to gather data on Soldiers from across all MOS who completed the TAPAS (and WPA and ICTL) at entry. These measures will again include JKTs, the ALQ, and supervisor ratings. Finally, the separation status of all Soldiers who took the TAPAS at entry is being tracked throughout the course of the research.

The May 2011 data file, which was the basis for analyses documented in this report, includes a total of 151,625 applicants who took the TAPAS. Of these total applicants, 141,483 were in the TOPS Applicant Sample. The Applicant Sample was determined by excluding Education Tier 3, AFQT Category V, and prior service applicants from the master data file. The validation sample sizes are considerably smaller, with the Schoolhouse Validation Sample comprising 4,976 Soldiers and the Validation Sample (which includes Soldiers for whom we only have administrative criterion data) comprising 46,188 Soldiers.

The JKT, ALQ, and administrative criterion measures exhibited acceptable and theoretically consistent psychometric properties. The Army-wide and MOS-specific PRS, however, continued to exhibit very low interrater reliability. The PRS instruments are currently being revised to change both content and format in an attempt to improve their psychometric characteristics. Details of these changes will be presented when we start including data from the new measures in the analysis data files. Until improvements can be implemented, results based on supervisor ratings should be interpreted with caution.

Summary of Evaluation Results to Date

TAPAS Construct Validity

The three versions of the TAPAS (13D-CAT, 15D-Static, and 15D-CAT) are consistent with one another in terms of their means, standard deviations, and patterns of intercorrelations (Allen et al., 2011). The two computer-adaptive versions of the TAPAS are particularly similar. Some of the TAPAS scales appear more similar across research and operational settings than others. The patterns

of relations between TAPAS scales and individual difference variables (AFQT scores, race, ethnicity, and gender), however, were generally consistent from the EEEM to TOPS settings. Keeping in mind that previous research has shown large differences between the experimental and operational use of temperament measures (White, Young, Hunter, & Rumsey, 2008), these results suggest that the use of the TAPAS in an operational setting is promising.

Validity for Soldier Selection

Consistent with previous analyses in the TOPS research program (Caramagno et al., 2011; Trippe, Caramagno, et al., 2011), results suggest that the TAPAS holds promise for predicting key criteria of interest. Incremental validity beyond the ASVAB is reasonably strong, especially for will-do criterion measures. This is despite the low reliability of the supervisor ratings.

Results of the composite formation analyses yielded alternative TOPS composites developed by HumRRO and DCG. Analyses conducted to evaluate these potential composites show that they outperform the previously developed composites in terms of predictive utility, and in an absolute sense, they would make a positive contribution to the Army's current selection system.

Potential for Soldier Classification

In the initial evaluation cycle, Trippe, Caramagno, et al. (2011) examined the classification potential of the TAPAS by looking at MOS differences in TAPAS score profiles. Mean differences (evaluated by computing the overall average root mean squared difference in scale scores) for the overall TAPAS were comparatively smaller than those observed in the ASVAB. The magnitude of the differences varied by TAPAS scale, however, often in ways that are consistent with a theoretical understanding of the scale and the MOS.

The classification results presented in this report, which are based on a stronger but still limited sample for purposes of classification analyses, further demonstrate that the TAPAS can provide incremental improvements beyond the ASVAB subtests for optimally assigning Soldiers to MOS. These incremental gains were observed in the dichotomous outcome variables (attrition and IMT restart) for which larger sample sizes in a greater number of MOS are available. In these cases, the reduction in predicted overall attrition or IMT restart is modest, but some MOS level results suggest a significant improvement. This is to some extent a result of the classification scenario modeled here, in which every Soldier must be assigned to an MOS according to the allocation percentages. Results from the JKT and ALQ criterion variables currently do not suggest that the TAPAS provides incremental improvements in classification beyond the ASVAB for the outcomes these measures represent. Nevertheless, these analyses currently suffer from a number of limitations related to the availability of criterion data within and across the MOS. We expect these analyses to become more informative as these criterion data continue to accumulate and we obtain JKT and ALQ data for additional MOS.

Results Summary

Taken together, evaluation results thus far suggest that, while the magnitude of the validity and classification coefficients are not as large as those found in the experimental EEEM research (Knapp & Heffner, 2010), the TAPAS holds promise for both selection and

classification-oriented purposes. Many of the scale-level coefficients are consistent with a theoretical understanding of the TAPAS scales, suggesting that the scales are measuring the characteristics that they are intended to measure. However, given the restricted nature of the matched criterion sample (in terms of sample characteristics) and the low reliability of the ratings data, these results should be considered preliminary.

Looking Ahead

Predictor Measures

MEPCOM will begin administering the WPA to Army applicants in CY2012 and so later evaluation cycles will include both WPA and ICTL as additional predictors. Soon, three new versions of TAPAS will also be introduced into the MEPS. Each of the 15-dimension versions will have nine core dimensions that are consistent across versions and include all of the scales in the "can-do" and "will-do" composites. Six dimensions, which were included in the original version of TAPAS and have shown promise for initial entry selection, are included on two of the three TAPAS versions. Six new scales are being tested and evaluated on a single TAPAS version (see Table 7.1). The dimensions will be evaluated for potential use as core dimensions on later versions of TAPAS. The current version of TAPAS will continue to be used in the research environment.

Table 7.1. TAPAS Dimensions Assessed

	Version A	Version B	Version C
Achievement	✓	✓	✓
Adjustment	✓	✓	✓
Adventure Seeking		✓	
Attention Seeking	✓	✓	✓
Commitment to Serve		✓	
Cooperation	✓	✓	
Courage			✓
Dominance	✓	✓	✓
Even Tempered	✓	✓	✓
Intellectual Efficiency	✓	✓	✓
Non-Delinquency	✓	✓	✓
Optimism	✓	✓	✓
Order	✓	✓	
Physical Conditioning	✓	✓	✓
Responsibility			✓
Self Control	✓		✓
Selflessness	✓	✓	
Sociability	✓		✓
Situational Awareness		✓	
Team Orientation			✓
Tolerance	✓		✓

Criterion Measures

In mid-2011, the MOS-specific and WBTD JKTs (both training and in-unit versions) were reviewed and updated with the assistance of Army subject matter experts. As part of this effort, additional items were added to the WBTD JKT in an effort to increase both its reliability and content representativeness. Additional items were also added to the 31B JKTs to cover content domains that have increased in relevance since the test blueprint was originally developed. In addition to updating and improving existing measures, we continued efforts to develop MOS-specific measures (both training and in-unit) for two occupations—Signal Support Specialist (25U) and Human Resources Specialist (42A).

We have also recently revised both the training and in-unit performance rating scales in an effort to improve their psychometric properties. For example, we have changed the format of the training MOS-specific rating scales to use a 5-point relative performance rating rather than a 7-point absolute performance rating and to greatly reduce the amount of reading required. The training Army-wide PRS have been similarly changed, and the number of dimensions rated has been reduced.

In-Unit Data Collections

Collection of data from Soldiers in units who took the TAPAS prior to enlistment began in April 2011. The data collection model closely mirrors that which was used in the Army Class research program (Knapp, Owens, et al., 2011). To the extent possible, we will visit major Army installations and reserve component training sites to collect Soldier and supervisor data in proctored settings. Other Soldiers will provide data from self-administered testing sessions.

Analyses

The semi-annual reports will include basic psychometric, validation, and incremental validation analyses. As needed, we will examine the comparability of new TAPAS versions to prior forms before determining if the data can be combined for purposes of analysis. Analysis strategies also will be developed to handle data produced by substantially revised performance rating scales which started being administered in fall 2011. This will be a particular challenge in the training validation sample and may require truncation of some future analyses to include only data provided by the newer measures to provide the best criterion-related validity evidence. Finally, the plan is to conduct classification-oriented analyses annually.

The next set of TOPS evaluation analyses will be conducted based on a data file constructed in December 2011. The sample sizes for this next evaluation are expected to be considerably larger, thus supporting additional analyses yielding more generalizable results.

REFERENCES

Allen, M.T., Cheng, Y.A., Putka, D.J., Hunter, A., & White L. (2010). Analysis and findings. In D.J. Knapp & T.S. Heffner (Eds.). *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Allen, M.T., Ingerick, M.J., & DeSimone, J.A. (2011). Psychometric evaluation of the TAPAS. In D.J. Knapp, T.S. Heffner, & L. White (Eds.) *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. *Psychometric Monograph* (No. 12).

Campbell, J.P., & Knapp, D.J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J.P., Hanson, M. A., & Oppler S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* Hillsdale, NJ: Erlbaum.

Caramagno, J.P., Allen, M. T., & Ingerick, M. J. (2011). Evidence for the predictive validity of the TAPAS. In D.J. Knapp & T.S. Heffner (Eds.), *Tier One Performance Screen initial operational test and evaluation: 2010 Annual Report* (FR 11-40). Alexandria, VA: Human Resources Research Organization.

Cassidy, T., & Lynn, R. (1989). A multifactorial approach to achievement motivation: The development of a comprehensive measure. *Journal of Occupational Psychology, 62,* 301-312.

Chernyshenko, O.S., & Stark, S. (October, 2007). *Criterion validity evidence for narrow temperament clusters: A meta-analysis of military studies*. Paper presented at the 49th annual conference of the International Military Testing Association. Gold Coast, AU.

Chernyshenko, O.S., Stark, S., Drasgow, F., & Roberts, B.W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88-106.

Chernyshenko, O.S., Stark, S., Woo, S., & Conz, G. (2008, April). *Openness to Experience: Its facet structure, measurement and validity*. Paper presented at at the 23nd annual conference for the Society of Industrial and Organizational Psychologists. New Orleans, LA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd edition*. Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S., (2003). *Applied multiple regression/ correlation analysis for the behavioral and social sciences* (3rd ed.), Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, M., Le, H., & Schantz, L. (2005). Job knowledge criterion tests. In D.J. Knapp & T.R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Technical Report 1168) (pp. 49-58). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

DeCorte, W. (2000). Estimating the classification efficiency of a test battery. *Educational and Psychological Measurement, 60,* 73-85.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Horst, P. (1954). A technique for the development of a differential predictor battery. *Psychometrika,* 68.

Horst, P. S. (1955). A technique for the development of an absolute prediction battery. *Psychometrika,* 69.

Ingerick, M., Diaz, T., & Putka, D. (2009). *Investigations into Army enlisted classification systems: Concurrent validation report* (Technical Report 1244). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35,* 1-19.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods, 7,* 238-257.

Kanfer, R., & Ackerman, P. L. (2000). Individual differences in work motivation: Further explorations of a trait framework. *Applied Psychology: An International Review, 49,* 470-482.

Knapp, D. J., & Campbell, R. C. (Eds). (2006). *Army enlisted personnel competency assessment program: Phase II report* (Technical Report 1174). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Heffner, T. S. (Eds.) (2009). *Predicting Future Force Performance (Army Class): End of Training Longitudinal Validation* (Technical Report 1257). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Heffner, T. S. (Eds.). (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Heffner, T. S. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: 2010 annual report* (Technical Report 1296). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., Heffner, T. S., & White, L. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., Owens, K. S., & Allen, M. T. (Eds.) (2011). *Validating future force performance measures (Army Class): First in-unit performance longitudinal validation* (Technical Report 1293). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D .J., & Tremble, T. R. (Eds) (2007). *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh*, *62,* 28–30.

Moriarty, K. O., Campbell, R. C., Heffner, T. S., & Knapp, D. J. (2009). *Validating future force performance measures (Army Class): Reclassification test and criterion development* (Research Product 2009-11). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Nord, R., & Schmitz, E. (1991). Estimating performance utility effects of alternative selection and classification policies. In J. Zeidner & C. D. Johnson (Eds.). *The economic benefit of predicting job performance: Vol 5. Estimating the gains of alternative policies* (pp.73-131), New York: Praeger.

Nye, C.D., Drasgow, F., Chernyshenko, O.S., Stark, S., Kubisiak, C., & White, L.A., *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as an MOS qualification instrument*. In preparation.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1990). *Numerical recipes: The art of scientific computing*. New York: Cambridge University Press.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*, 959-981.

Putka, D. J. & Van Iddekinge, C. H. (2007). Work Preferences Survey. In D. J. Knapp & T. R. Tremble (Eds.), *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Technical Report 1205). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99-115.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3-32.

Rosse, R. L., Campbell, J. P., & Peterson, N. G. (2001). Personnel classification and differential job assignments: Estimating classification gains. In J. P Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 453-506). Mahwah, NJ: Lawrence Erlbaum Associates.

Schmitt, N., & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. *Journal of Applied Psychology, 84*(1), 50-57.

Scholarios, D., Johnson, C. D., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 79,* 412-424.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment* [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Stark, S. E., Hulin, C. L., Drasgow, F., & Lopez-Rivas, G. (2006). *Technical report 2 for the SBIR Phase II funding round, topic A04-029: Behavior domains assessed by TAPAS* (DCG200608). Urbana, IL: Drasgow Consulting Group.

Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2010a). Adaptive testing with the multi-unidimensional pairwise preference model. Manuscript submitted for publication.

Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2010b). Tailored adaptive personality assessment system (TAPAS-95s). In D. J. Knapp & T. S. Heffner (Eds.) *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (September, 2010c).*Update on the Tailored Adaptive Personality Assessment System (TAPAS): Results and ideas to meet the challenges of high stakes testing.* Paper presented at the 52nd annual conference of the International Military Testing Association. Lucerne, Switzerland.

Stark, S. E., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise preference model. *Applied Psychological Measurement, 29,* 184-201.

Stark, S. E., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91,* 25-39.

Tonidandel, S. & LeBreton, J. M. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weight analysis. *Organizational Research Methods, 13*, 767-781.

Trippe, D. M., Caramagno, J. P., Allen, M. T., & Ingerick, M. J. (2011). Initial evidence for the predictive validity and classification potential of the TAPAS. In D. J. Knapp, T. S. Heffner, & L. White (Eds.), *Tier One Performance Screen initial operational test and evaluation: Early results* (TR 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Trippe, D. M., Ford, F., Bynum, B., & Moriarty, K. O. (2011). Data file development. In D. J. Knapp & T. S. Heffner (Eds.), *Tier Once Performance Screen initial operational test and evaluation: 2010 annual report* (FR 11-40). Alexandria, VA: Human Resources Research Organization.

Van Iddekinge, C. H., Putka, D .J., & Sager, C. E. (2005). Attitudinal criteria. In D. J. Knapp & T. R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (pp. 89-104) (Technical Report 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

White, L. A., Young, M. C., Heggestad, E. D., Stark, S., Drasgow, F., & Piskator, G. (2004). *Development of a non-high school diploma graduate pre-enlistment screening model to enhance the future force*. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(3), 291-295.

APPENDIX A

PREDICTOR MEASURE PSYCHOMETRIC PROPERTIES

Table A.1. Raw Means and Standard Deviations for the TOPS IOT&E TAPAS Scales by Version

	TAPAS Version					
	13D-CAT[a] (n = 1,395)		15D-Static (n = 12,899)		15D-CAT (n = 121,614)	
TAPAS Composite/Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*
Individual Composite/Scale						
Achievement	0.236	0.493	0.274	0.500	0.154	0.484
Adjustment	--	--	0.158	0.586	-0.007	0.573
Attention Seeking	-0.222	0.555	-0.259	0.532	-0.196	0.521
Cooperation	0.028	0.392	-0.070	0.391	-0.065	0.380
Dominance	0.068	0.598	-0.025	0.588	0.030	0.576
Even Tempered	0.132	0.515	0.257	0.484	0.150	0.467
Generosity	-0.172	0.429	-0.191	0.445	-0.195	0.441
Intellectual Efficiency	0.100	0.604	-0.101	0.588	-0.021	0.583
Non-Delinquency	0.103	0.462	0.120	0.455	0.077	0.464
Optimism	0.175	0.463	0.276	0.506	0.127	0.450
Order	-0.411	0.564	-0.398	0.575	-0.405	0.540
Physical Conditioning	-0.019	0.616	-0.048	0.618	0.037	0.613
Self Control	--	--	0.095	0.528	0.041	0.540
Sociability	-0.029	0.620	-0.213	0.593	-0.042	0.581
Tolerance	-0.240	0.591	-0.263	0.586	-0.219	0.557
Original TAPAS Composites						
TAPAS Can-Do Composite	0.006	2.712	-0.030	2.669	-0.057	2.717
TAPAS Will-Do Composite	0.005	2.460	-0.039	2.365	-0.014	2.381

Note. Results are limited to the Applicant Sample (Non-prior service, Education Tier 1 and 2, AFQT Category IV and above).
[a]This version of the TAPAS is no longer being administered.

Table A.2. TAPAS Facet Scale and AFQT Intercorrelations

TAPAS Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Achievement																	
2. Adjustment	**.11**																
3. Attention Seeking	**.05**	**.11**															
4. Cooperation	**.10**	**.11**	**.05**														
5. Dominance	**.32**	**.10**	**.20**	.00													
6. Even Tempered	**.10**	**.18**	**-.01**	**.23**	**-.06**												
7. Generosity	**.09**	**-.02**	**-.07**	**.22**	.00	**.10**											
8. Intellectual Efficiency	**.26**	**.19**	**.08**	**.02**	**.24**	**.08**	**-.02**										
9. Non-Delinquency	**.17**	.00	**-.13**	**.16**	**-.01**	**.18**	**.12**	**.01**									
10. Optimism	**.19**	**.27**	**.16**	**.15**	**.17**	**.18**	**.03**	**.10**	**.09**								
11. Order	**.15**	**-.07**	**-.08**	.00	**.05**	**-.04**	**.05**	**.02**	**.07**	**-.02**							
12. Physical Conditioning	**.15**	**.07**	**.12**	**-.01**	**.18**	**-.08**	**-.04**	**.05**	**-.02**	**.10**	**.03**						
13. Self Control	**.20**	**.05**	**-.13**	**.12**	**.03**	**.20**	**.06**	**.15**	**.26**	**.07**	**.15**	**-.05**					
14. Sociability	**.05**	**.11**	**.36**	**.17**	**.22**	**.03**	**.06**	.00	**-.05**	**.23**	**-.04**	**.13**	**-.12**				
15. Tolerance	**.11**	**.02**	**.04**	**.14**	**.06**	**.12**	**.30**	**.07**	**.05**	**.08**	**.04**	**-.06**	**.09**	**.11**			
16. AFQT	**.09**	**.09**	**.10**	**.01**	**.08**	**.08**	**-.06**	**.41**	.00	**.02**	**-.17**	**.04**	**-.01**	**-.08**	**-.01**		
17. TAPAS Will-Do Composite	**.56**	**.10**	**-.40**	**.17**	**.10**	**.49**	**.14**	**.13**	**.60**	**.16**	**.12**	**.38**	**.30**	**-.08**	**.08**	**.04**	
18. TAPAS Can-Do Composite	**.62**	**.27**	**.05**	**.24**	**.24**	**.55**	**.12**	**.52**	**.52**	**.55**	**.07**	**.07**	**.32**	**.09**	**.16**	**.21**	**.71**

Note. N = 134,513-135,908. Coefficients in bold are statistically significant, $p < .05$. Results are limited to the Applicant Sample (Non-prior service, Education Tier 1 and 2, AFQT Category IV and above).

Table A.3. Descriptive Statistics for the ASVAB

Measure/Scale	*n*	*M*	*SD*	*Min*	*Max*
AFQT	141,483	57.24	23.39	10	99
ASVAB Subtests					
General Science (GS)	140,388	51.71	8.48	19	76
Arithmetic Reasoning (AR)	140,388	52.63	7.78	18	72
Word Knowledge (WK)	140,388	51.37	8.20	16	76
Paragraph Comprehension (PC)	140,388	52.79	7.18	21	69
Math Knowledge (MK)	140,388	53.40	7.09	24	73
Electronics Information (EI)	140,388	52.14	9.15	16	84
Auto and Shop Information (AS)	140,388	50.26	9.47	19	86
Mechanical Comprehension (MC)	140,388	53.56	8.49	14	82
Assembling Objects (AO)	139,767	55.09	7.89	25	70
ASVAB Aptitude Area Composites					
Clerical (CL)	140,396	105.93	14.19	35	152
Combat (CO)	140,396	105.87	15.10	29	160
Electronics (EL)	140,396	105.66	15.10	29	160
Field Artillery (FA)	140,396	106.03	15.02	28	159
General Maintenance (GM)	140,396	105.42	15.54	28	161
Mechanical Maintenance (MM)	140,396	104.65	16.51	25	165
Operators and Food Service (OF)	140,396	105.41	15.51	27	160
Signal Communication (SC)	140,396	106.01	14.74	29	159
Skill Technical (ST)	140,396	105.86	14.76	32	157

Note. Results are limited to the Applicant Sample (non-prior service, Education Tier 1 and 2, AFQT Category IV and above).

Table A.4. ASVAB Subset and AFQT Intercorrelations

ASVAB Subtests		1	2	3	4	5	6	7	8	9
1	General Science (GS)									
2	Arithmetic Reasoning (AR)	**.56**								
3	Word Knowledge (WK)	**.73**	**.49**							
4	Paragraph Comprehension (PC)	**.67**	**.57**	**.72**						
5	Math Knowledge (MK)	**.45**	**.70**	**.36**	**.44**					
6	Electronics Information (EI)	**.69**	**.48**	**.60**	**.55**	**.30**				
7	Auto and Shop Information (AS)	**.51**	**.31**	**.40**	**.36**	**.06**	**.69**			
8	Mechanical Comprehension (MC)	**.68**	**.61**	**.55**	**.57**	**.42**	**.70**	**.62**		
9	Assembling Objects (AO)	**.37**	**.48**	**.29**	**.36**	**.39**	**.36**	**.27**	**.53**	
10	AFQT	**.75**	**.82**	**.82**	**.81**	**.73**	**.60**	**.37**	**.66**	**.45**

Note. N = 139,767-140,388. Coefficients in bold are statistically significant, $p < .05$. Results are limited to the Applicant Sample (Non-prior service, Education Tier 1 and 2, AFQT Category IV and above).

Table A.5. Race, Ethnic, and Gender Subgroup Means and Standard Deviations

	Mean						Standard Deviation					
Predictor	All	Male	Female	White	Black	Hispanic	All	Male	Female	White	Black	Hispanic
AFQT	57.24	58.35	52.67	59.44	45.48	47.14	23.39	23.49	22.40	23.06	20.75	21.41
TAPAS Scales												
Achievement	--[a]	-.01	.02	.02	-.05	-.06	--	1.01	0.96	1.01	0.96	0.96
Adjustment	--	.06	-.23	.02	-.06	-.11	--	1.00	0.98	1.01	0.98	0.95
Attention Seeking	--	.02	-.09	.02	-.07	-.03	--	1.00	0.99	1.01	0.94	0.95
Cooperation	--	.00	-.01	-.01	.01	-.03	--	1.00	0.99	1.00	0.98	0.98
Dominance	--	.02	-.10	.01	.03	.00	--	1.01	0.97	1.02	0.90	0.94
Even Tempered	--	.02	-.08	.00	.00	-.07	--	1.00	1.01	1.00	0.99	0.95
Generosity	--	-.08	.31	-.02	.11	.05	--	0.99	0.98	1.01	0.98	0.97
Intellectual Efficiency	--	.04	-.15	.02	-.09	-.13	--	1.01	0.94	1.01	0.92	0.93
Non-Delinquency	--	-.03	.12	.00	.04	-.05	--	1.01	0.97	1.00	0.98	0.96
Optimism	--	.01	-.04	.01	.03	.00	--	1.00	1.01	1.00	0.97	0.95
Order	--	-.03	.12	-.05	.18	.13	--	0.99	1.03	1.00	0.96	0.96
Physical Conditioning	--	.08	-.31	.04	-.16	-.07	--	0.99	0.97	1.01	0.95	0.94
Self Control	--	.00	.01	-.03	.17	.07	--	1.00	1.01	1.00	0.99	0.99
Sociability	--	.00	.01	.01	-.05	.00	--	1.00	1.00	1.01	0.94	0.95
Tolerance	--	-.07	.27	-.05	.19	.22	--	1.00	0.96	1.01	0.92	0.90
Original TAPAS Composites												
Can-Do	--	.01	-.04	.02	-.02	-.11	--	1.00	1.00	1.00	0.98	0.97
Will-Do	--	.02	-.06	.02	-.04	-.09	--	1.00	1.00	1.00	0.98	0.97

Note. All *n* = 134,513-141,483; Male *n* = 107,937-113,875; Female *n* = 26,511-27,540; White *n* = 98,186-102,819; Black *n* = 19,358-20,722; Hispanic *n* = 20,009-20,802. White and Black include both Hispanic and non-Hispanic. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. Results are limited to non-prior service, Education Tier 1 and 2, AFQT Category IV and above applicants.

[a]Values for the TAPAS scales are omitted because they had been standardized to a mean of 0 and a standard deviation of 1.

Table A.6. Race, Ethnic, and Gender Subgroup Differences in Means

	Mean Differences[a] (Cohen's *d*)		
Predictor	Female	Black	Hispanic
AFQT	-.24 ***	-.61 ***	-.52 ***
TAPAS Scales			
Achievement	.03	-.06 ***	-.07 ***
Adjustment	-.29 ***	-.08 ***	-.13 ***
Attention-Seeking	-.12 ***	-.09 ***	-.03 ***
Cooperation	-.02 *	.01	-.03 ***
Dominance	-.12 ***	.02	.01
Even Tempered	-.10 ***	.00	-.09 ***
Generosity	.39	.13	.05
Intellectual Eff.	-.19 ***	-.11 ***	-.14 ***
Non-Delinquency	.15	.03	-.06 ***
Optimism	-.05 ***	.02	.00
Order	.15	.23	.15
Physical Cond.	-.39 ***	-.20 ***	-.08 ***
Self-Control	.01	.20	.09
Sociability	.01	-.06 ***	.00
Tolerance	.34	.24	.26
Original TAPAS Composites			
Can-Do	-.06 ***	-.05 ***	-.13 ***
Will-Do	-.05 ***	-.08 ***	-.11 ***

Note. White includes both Hispanic and non-Hispanic; Black includes both Hispanic and non-Hispanic. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. Results are limited to non-prior service, Education Tier 1 and 2, AFQT Category IV and above applicants.

[a] Each value in the three columns on the right, labeled *Difference in Means*, represents the difference between the mean score for the minority group (i.e., Female, Black, Hispanic) and the mean score for the reference group (i.e., Male, White) in terms of Cohen's *d*. A negative value indicates that the minority group's mean is less than the referent group's mean. Significance differences are based on independent samples *t*-test analyses of mean differences between the two groups of interest. Significant differences in means are asterisked only where the difference favors the referent group and *$p < .05$, ** $p < .01$, *** $p < .001$.

APPENDIX B

CRITERION MEASURE PSYCHOMETRIC PROPERTIES IN FULL SCHOOLHOUSE SAMPLE

Table B.1. Descriptive Statistics for Training Criteria Based on the Full Schoolhouse Sample

Measure/Scale	*n*	*M*	*SD*	*Min*	*Max*	*α*	*IRR*
Army Life Questionnaire (ALQ)							
Affective Commitment[a]	24,486	3.85	0.68	1.00	5.00	.86	--
Attrition Cognition[b]	24,486	1.54	0.62	1.00	5.00	.78	--
Army Life Adjustment[a]	24,486	4.05	0.66	1.00	5.00	.86	--
MOS Fit[a]	24,486	3.79	0.85	1.00	5.00	.93	--
Army Fit[a]	24,486	4.05	0.60	1.00	5.00	.86	--
Training Achievement[c]	24,454	0.39	0.60	0.00	2.00	--	--
Training Restarts[c]	24,485	0.40	0.64	0.00	4.00	--	--
Disciplinary Incidents[c]	12,308	0.25	0.59	0.00	7.00	--	--
Last APFT Score	24,189	248.36	32.36	5.00	300.00	--	--
MOS-Specific Job Knowledge Test (JKT)							
11B/11C/11X/18X	7,488	61.94	9.78	20.93	88.37	.75	--
19K	116	61.91	10.05	30.00	82.00	.75	--
31B	3,089	70.17	8.61	34.95	93.20	.78	--
68W	5,039	74.52	9.77	30.43	96.74	.85	--
88M	2,537	66.28	11.15	33.33	94.44	.78	--
91B	849	57.48	13.69	23.71	88.66	.90	--
WTBD Job Knowledge	23,748	65.60	12.62	9.68	100.00	.64	--
Army-Wide Performance Rating Scales[d]							
Can-Do	8,808	4.91	1.14	1.00	7.00	.89	.11
Commitment and Adjustment	9,109	5.02	1.22	1.00	7.00	--	.17
Effort and Discipline	9,135	4.93	1.16	1.00	7.00	.84	.22
Physical Fitness and Bearing	9,059	4.83	1.22	1.00	7.00	--	.24
Work with Others	9,079	4.85	1.16	1.00	7.00	.83	.19
Overall Performance	8,950	3.50	0.83	1.00	5.00	--	.32
MOS-Specific Performance Rating Composite Scores							
Total (combined across MOS)	7,626	4.72	0.96	1.00	7.00	--	--
11B/11C/11X/18X	3,246	4.86	0.92	1.00	7.00	.94	.18
19K	70	5.07	0.61	2.71	6.86	.87	.52
31B	1,059	4.96	0.99	1.00	7.00	.95	.17
68W	2,400	4.41	0.86	1.00	7.00	.93	.04
88M	671	4.79	0.94	2.00	7.00	.93	.00
91B	180	4.57	1.65	1.00	7.00	.97	.21

Note. Job knowledge scores are percent correct. WTBD = Warrior Tasks and Battle Drills. IRR = Interrater Reliability computed using G(q,k) (Putka, Le, McCloy, & Diaz, 2008).

[a] These items were responded to using agreement scales (1=Strongly Disagree to 5=Strongly Agree).

[b] This construct was measured by items using agreement scale (same as above) and often scale (1=Never to 5=Very Often).

[c] These scales are the total number of 'YES' responses to a series of yes/no questions about things that happened in training.

[d] The possible Army-wide and MOS-Specific Performance Rating Composite Scores are between 1 and 7, except for overall performance which ranges from 1 to 5.

Table B.2. Descriptive Statistics for Schoolhouse Criteria by MOS from the Full Schoolhouse Sample

	11B		19K		25U		31B		42A		68W		88M		91B	
Measure/Scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Army Life Questionnaire (ALQ)																
Affective Commitment	3.89	0.67	3.91	0.67	3.58	0.72	3.94	0.63	3.87	0.67	3.73	0.72	3.94	0.65	3.78	0.70
Attrition Cognition	1.52	0.61	1.49	0.64	1.74	0.73	1.51	0.58	1.59	0.62	1.58	0.62	1.49	0.60	1.66	0.68
Adjust to Army Life	4.05	0.68	3.89	0.70	3.99	0.62	4.07	0.63	4.00	0.73	4.00	0.65	4.11	0.63	4.03	0.69
MOS Fit	3.85	0.80	3.26	0.93	3.34	0.84	3.89	0.79	3.55	0.90	3.96	0.83	3.31	0.85	3.61	0.89
Army Fit	4.07	0.59	4.12	0.61	3.84	0.63	4.12	0.54	4.08	0.61	3.93	0.62	4.14	0.59	3.96	0.63
Training Achievement	0.45	0.67	0.31	0.59	0.46	0.61	0.35	0.58	0.48	0.65	0.30	0.49	0.40	0.58	0.39	0.56
Training Restarts	0.30	0.55	0.34	0.56	0.73	0.81	0.26	0.51	0.72	0.80	0.56	0.73	0.48	0.68	0.54	0.71
Disciplinary Incidents	0.24	0.57	0.26	0.55	--	--	0.27	0.62	--	--	0.36	0.76	0.35	0.66	0.28	0.66
Last APFT Score	245.28	32.83	252.96	26.45	243.80	34.38	256.04	30.54	243.42	33.49	250.00	31.55	245.36	32.11	243.18	30.32
MOS-Specific JKT	61.74	9.77	61.91	10.05	--	--	70.17	8.61	--	--	74.52	9.77	66.28	11.15	57.48	13.69
WTBD JKT	64.78	12.52	67.44	11.63	58.97	12.45	69.94	10.52	58.66	13.18	68.42	11.38	62.03	13.12	58.20	12.84
Army-Wide PRS																
Can Do	4.89	1.06	5.69	0.83	4.37	1.16	5.37	1.15	--	--	4.77	1.16	4.73	1.01	4.80	1.48
Commitment and Adjustment	5.09	1.19	5.40	0.95	4.39	1.22	5.29	1.30	--	--	4.90	1.14	4.90	1.09	4.81	1.78
Effort and Discipline	4.93	1.18	5.45	0.85	4.41	1.07	5.22	1.22	--	--	4.89	1.09	4.81	1.02	4.64	1.60
Physical Fitness and Bearing	4.82	1.22	5.43	1.20	4.38	1.26	5.01	1.26	--	--	4.80	1.17	4.67	1.08	4.56	1.73
Work with Others	4.81	1.19	5.32	1.11	4.43	1.23	5.11	1.25	--	--	4.81	1.06	4.78	0.99	4.91	1.47
Overall Performance	3.44	0.84	3.60	0.92	3.28	0.82	3.55	0.92	--	--	3.52	0.79	3.64	0.71	3.37	1.08
MOS-Specific Performance Composite	4.87	0.93	5.07	0.61	--	--	4.96	0.99	--	--	4.41	0.86	4.79	0.92	4.58	1.62

Note. MOS-specific JKT and WTBD JKT test scores are percent correct.

Table B.3. Army Life Questionnaire (ALQ) Intercorrelations for Full Schoolhouse Sample

Scale	1	2	3	4	5	6	7	8
1. Affective Commitment								
2. Attrition Cognitions	**-.62**							
3. Adjust to Army Life	**.46**	**-.55**						
4. MOS Fit	**.48**	**-.42**	**.36**					
5. Army Fit	**.83**	**-.68**	**.62**	**.49**				
6. Training Achievement	**.07**	**-.05**	**.13**	**.05**	**.08**			
7. Training Restarts	**-.07**	**.13**	**-.21**	**-.09**	**-.10**	**-.11**		
8. Disciplinary Incidents	**-.09**	**.14**	**-.19**	**-.10**	**-.13**	**-.07**	**.19**	
9. Last APFT Score	**.05**	**-.12**	**.24**	**.09**	**.11**	**.23**	**-.28**	**-.16**

Note. All correlation are statistically significant ($p < .05$). N = 12,167–24,486.

Table B.4. Performance Rating Scales (PRS) Intercorrelations for Full Schoolhouse Sample

Scale	1	2	3	4	5	6
Army-Wide Performance Rating Scales (PRS)						
1. Can Do						
2. Commitment and Adjustment	**.75**					
3. Effort and Discipline	**.72**	**.81**				
4. Physical Fitness and Bearing	**.65**	**.69**	**.73**			
5. Work with Others	**.78**	**.78**	**.79**	**.68**		
6. Overall Performance	**.55**	**.57**	**.61**	**.58**	**.61**	
MOS-Specific Performance Ratings Composites						
7. Combined MOS-Specific PRS	**.72**	**.63**	**.61**	**.55**	**.63**	**.49**
8. 11B	**.74**	**.66**	**.66**	**.60**	**.68**	**.54**
9. 19K	**.79**	**.69**	**.73**	**.70**	**.66**	**.72**
10. 31B	**.77**	**.64**	**.65**	**.57**	**.67**	**.56**
11. 68W	**.58**	**.47**	**.44**	**.37**	**.45**	**.34**
12. 88M	**.73**	**.66**	**.60**	**.60**	**.65**	**.54**
13. 91B	**.93**	**.83**	**.83**	**.76**	**.81**	**.66**

Note. All correlations are statistically significant ($p < .05$). AW PRS intercorrelations n = 8,639-9,135. 11B n = 2,971-2,974, 19K n = 67, 31B n = 969-982, 68W n = 1,474-1,699, 88M n = 606-623, 91B n = 154-173.

Table B.5. Correlations between the Army Life Questionnaire (ALQ) and Job Knowledge Tests (JKT) in Full Schoolhouse Sample

	Job Knowledge Tests							
ALQ Scales	Combined	11B	19K	31B	68W	88M	91B	WTBD JKT
Affective Commitment	**.07**	**.09**	-.06	**.06**	**.06**	.03	**.09**	**.08**
Attrition Cognition	**-.14**	**-.16**	-.05	**-.12**	**-.15**	**-.10**	**-.15**	**-.16**
Adjust to Army Life	**.12**	**.12**	-.08	**.13**	**.12**	**.11**	**.14**	**.14**
MOS Fit	**.09**	**.10**	-.10	**.04**	**.14**	.04	**.23**	**.13**
Army Fit	**.12**	**.14**	-.06	**.09**	**.12**	**.07**	**.13**	**.13**
Training Achievement	**-.10**	**-.14**	**-.23**	**-.05**	-.01	**-.15**	**-.11**	**-.08**
Training Restarts	**-.08**	**-.05**	-.01	**-.11**	**-.07**	**-.14**	**-.07**	**-.12**
Disciplinary Incidents	**-.04**	**-.04**	.04	-.01	-.05	-.03	-.04	**-.06**
Last APFT Score	.01	**.04**	-.05	.02	.01	-.02	**-.09**	**.08**

Note. WTBD = Warrior Tasks and Battle Drills. Combined = MOS-specific JKT scores combined into one variable. Significant correlation coefficients are bolded ($p < .05$). Combined n = 9,628-18,951; 11B n =7,332-7,403; 19K n =116; 31B n= 720-3,068 68W n = 783-4,999; 88M n = 469-2,519; 91B n = 137-846; WTBD n = 11,825-23,514.

Table B.6. Correlations between the Army Life Questionnaire (ALQ) and Performance Rating Scales (PRS) in Full Schoolhouse Sample

	Army Life Questionnaire (ALQ)								
Performance Rating Scales	AFF COM	ATT COG	LIFE ADJ	MOS Fit	Army Fit	TRN ACH	TRN RST	DSC INC	LAST APFT
Army-Wide Performance Rating Scales									
Can Do	**.08**	**-.09**	**.12**	**.09**	**.09**	**.05**	**-.12**	**-.12**	**.13**
Commitment and Adjustment	**.08**	**-.09**	**.12**	**.08**	**.10**	**.07**	**-.11**	**-.14**	**.14**
Effort and Discipline	**.09**	**-.10**	**.12**	**.09**	**.10**	**.07**	**-.10**	**-.16**	**.14**
Physical Fitness and Bearing	**.07**	**-.11**	**.16**	**.10**	**.10**	**.10**	**-.15**	**-.12**	**.27**
Work with Others	**.07**	**-.09**	**.12**	**.08**	**.09**	**.08**	**-.10**	**-.13**	**.16**
Overall Performance	**.07**	**-.12**	**.17**	**.09**	**.11**	**.14**	**-.14**	**-.15**	**.23**
MOS-Specific Performance Ratings Composite									
Combined MOS-Specific PRS	**.08**	**-.09**	**.12**	**.05**	**.10**	**.09**	**-.11**	**-.12**	**.10**
11B	**.09**	**-.11**	**.14**	**.12**	**.10**	**.08**	**-.10**	**-.12**	**.17**
19K	.11	-.20	**.31**	.02	.17	.15	**-.34**	**-.26**	**.41**
31B	**.09**	**-.15**	**.18**	**.08**	**.14**	**.13**	**-.13**	-.09	**.10**
68W	.01	-.03	**.06**	.03	.02	**.05**	**-.07**	-.07	**.04**
88M	.00	.03	.02	.00	.01	.02	.00	-.03	**.09**
91B	.05	.01	.09	.03	.01	.01	.03	--	-.08

Note. Significant correlation coefficients are bolded ($p < .05$). AW PRS N = 4,093-8,888. MOS-Specific PRS Combined n = 4,093-7,438; 11B n =3,109-3,139; 19K n = 70, 31B n = 428-1,047, 68W n = 192-2,374, 88M n = 265-629, 91B n = 176-179. AFFCOM=Affective Commitment; ATTCOG=Attrition Cognition; LIFEADJ=Adjust to Army Life; TRNACH=Training Achievement; TRNRST=Training Restart; DSCINC=Disciplinary Incidents; LASTAPFT=Last APFT Score.

Table B.7. Correlations between Job Knowledge Tests (JKT) and Performance Rating Scales (PRS) in Full Schoolhouse Sample

	MOS-Specific Job Knowledge Test (JKT)							
Performance Rating Scales	Total	11B	19K	31B	68W	88M	91B	WTBD JKT
Army-Wide Performance Rating Scales								
Can Do	.01	**.04**	.18	**.11**	**-.07**	-.01	.01	**.08**
Commitment and Adjustment	**.03**	**.04**	.09	**.10**	-.02	.03	.00	**.09**
Effort and Discipline	**.06**	**.04**	**.23**	**.13**	.01	**.09**	.04	**.11**
Physical Fitness and Bearing	**.04**	**.06**	.11	**.14**	-.01	.07	-.06	**.09**
Work with Others	**.03**	**.07**	.11	**.11**	-.05	.07	-.02	**.09**
Overall Performance	**.06**	.03	.20	**.15**	.01	.00	.01	**.11**
MOS-Specific Performance Ratings Composite								
Combined MOS-Specific PRS	**.03**	.03	.23	**.08**	.02	.01	.01	**.05**
11B	.03	.03	--	--	--	--	--	**.06**
19K	.23	--	.23	--	--	--	--	**.38**
31B	**.08**	--	--	**.08**	--	--	--	**.13**
68W	.02	--	--	--	.02	--	--	.02
88M	.01	--	--	--	--	.01	--	.05
91B	.01	--	--	--	--	--	.01	.09

Note. Significant correlation coefficients are bolded ($p < .05$). AW PRS N = 116-7,292. MOS-Specific PRS Combined n = 116-7,243; 11B n =2,457-3,062; 19K n = 61-66, 31B n = 942-1,035, 68W n = 2,174-2,320, 88M n = 510-581, 91B n = 116-179.

Table B.8. Descriptive Statistics for Administrative Criteria Based on the TOPS Validation Sample by MOS

	11B/11C/11X/18X			*19K*			*25U*			*31B*		
Administrative Criterion	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$
Attrtion[a]												
3-Month Cumulative	6,004	472	7.9	454	19	4.2	344	24	7.0	501	30	6.0
6-Month Cumulative	4,889	589	12.0	307	35	11.4	216	18	8.3	275	33	12.0
9-Month Cumulative	2,614	378	14.5	152	19	12.5	65	11	16.9	159	25	15.7
Initial Military Training (IMT) Criteria	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$
Restarted at Least Once During IMT	4,467	711	15.9	186	32	17.2	128	4	3.1	471	62	13.2
Restarted at Least Once During IMT for Pejorative Reasons	4,440	684	15.4	183	29	15.8	128	4	3.1	462	53	11.5
Restarted at Least Once During IMT for Academic Reasons	4,069	313	7.7	167	13	7.8	127	3	2.4	434	25	5.8
AIT School Grades	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*
Overall Average (Unstandardized)	--	--	--	--	--	--	170	93.04	3.50	--	--	--
Overall Average (Standardized within MOS)	--	--	--	--	--	--	170	.05	0.95	--	--	--

	42A			*68W*			*88M*			*91B*		
Administrative Criterion	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$	N [b]	N_{Attrit}	$\%_{Attrit}$
Attrtion[a]												
3-Month Cumulative	233	12	5.2	1,405	79	5.6	871	59	6.8	853	63	7.4
6-Month Cumulative	176	17	9.7	1,014	80	7.9	699	78	11.2	677	72	10.6
9-Month Cumulative	90	10	11.1	605	57	9.4	486	54	11.1	325	46	14.2
Initial Military Training (IMT) Criteria	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$	N [c]	$N_{Restart}$	$\%_{Restart}$
Restarted at Least Once During IMT	344	19	5.5	620	120	19.4	1,151	199	17.3	594	61	10.3
Restarted at Least Once During IMT for Pejorative Reasons	343	18	5.2	589	89	15.1	1,021	68	6.7	587	53	9.0
Restarted at Least Once During IMT for Academic Reasons	341	16	4.7	620	120	19.4	1,148	196	17.1	583	50	8.6
AIT School Grades	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*	N [d]	*M*	*SD*
Overall Average (Unstandardized)	--	--	--	91	86.37	8.95	--	--	--	--	--	--
Overall Average (Standardized within MOS)	--	--	--	90	0.10	0.88	--	--	--	--	--	--

Note. Results are limited to non-prior service, Education Tier 1 and 2, AFQT Category IV or higher Soldiers.

[a] Attrition results reflect Regular Army Soldiers only. Note that attrition estimates are more unstable as the sample sizes drop, which explains why 9-month attrition can be lower than 6-month attrition for some MOS even though the months are cumulative.

[b] N = number of Soldiers with attrition data at the time data were extracted. N_{Attrit} = number of Soldiers who attrited. $\%_{Attrit}$ = percentage of Soldiers who attrited [(N_{Attrit} /N) x 100].

[c] N = number of Soldiers with ATRRS IMT data at the time data were extracted. $N_{Restart}$ = number of Soldiers who restarted at least once during IMT. $\%_{Restart}$ = percentage of Soldiers who restarted at least once during IMT [($N_{Restart}$ /N) x 100].

[d] N = number of Soldiers with RITMS AIT school grade data. Standardized school grades were not computed for MOS with insufficient sample size (n < 15).

Table B.9. Correlations between the Army Life Questionnaire (ALQ) Criteria and Administrative Criteria

	Army Life Questionnaire (ALQ)								
Administrative Criterion	1	2	3	4	5	6	7	8	9
Initial Military Training (IMT) Criteria									
Restarted at Least Once During IMT	.03	**-.06**	**.04**	**.07**	.02	.02	**-.18**	**-.11**	**.07**
Restarted at Least Once During IMT for Pejorative Reasons	.03	**-.07**	**.06**	**.05**	.02	.01	**-.18**	**-.10**	**.08**
Restarted at Least Once During IMT for Academic Reasons	.02	**-.04**	.02	**.07**	.01	.01	**-.17**	**-.08**	**.05**
AIT School Grades									
Overall Average (Unstandardized)	-.02	-.05	-.09	.03	.04	.04	.16	--	**-.24**
Overall Average (Standardized within MOS)	-.12	-.02	.02	-.07	-.05	-.01	.04	--	-.18

Note. Significant correlation coefficients are bolded ($p < .05$). IMT Criteria n = 1,561-2,833, AIT School Grades Unstandardized n = 86-87, Standardized n = 85-86. Attrition is excluded because individuals that attrit in their first months are not eligible to take the research-only assessments.1=Affective Commitment; 2=Attrition Cognitions; 3=Adjust to Army Life; 4=MOS Fit; 5=Army Fit; 6=Training Achievement; 7=Training Restarts; 8=Disciplinary Incidents; 9=Last APFT Score

Table B.10. Correlations between the Performance Rating Scales (PRS) and Administrative Criteria

Administrative Criterion	Army-Wide Performance Rating Scales (PRS) 1	2	3	4	5	6	MOS-Specific PRS
Initial Military Training (IMT) Criteria							
Restarted at Least Once During IMT	-.01	.06	.00	.01	.01	-.04	-.04
Restarted at Least Once During IMT for Pejorative Reasons	.02	**.07**	.03	.04	.01	.03	.03
Restarted at Least Once During IMT for Academic Reasons	-.02	.04	-.02	-.02	.02	-.05	-.05
AIT School Grades							
Overall Average (Unstandardized)	-.05	-.03	.09	.06	-.02	-.22	-.22
Overall Average (Standardized within MOS)	-.02	-.02	.10	.07	-.02	-.20	-.20

Note. Significant correlation coefficients are bolded ($p < .05$). IMT Criteria n = 770-873, AIT School Grades n = 38-43. Attrition is excluded because individuals that attrit in their first months are not eligible to take the research-only assessments.1=Can Do; 2=Commitment and Adjustment; 3=Effort and Discipline; 4=Physical Fitness and Bearing; 5=Work with Others; 6=Overall Performance.

Table B.11. Correlations between Research Only and Administrative Criteria

Administrative Criterion	MOS-Specific JKT	WTBD JKT
Initial Military Training (IMT) Criteria		
Restarted at Least Once During IMT	**.05**	**.04**
Restarted at Least Once During IMT for Pejorative Reasons	.03	.03
Restarted at Least Once During IMT for Academic Reasons	**.05**	**.04**
AIT School Grades		
Overall Average (Unstandardized)	**.34**	**.33**
Overall Average (Standardized within MOS)	**.32**	**.40**

Note. Significant correlation coefficients are bolded ($p < .05$). IMT Criteria n = 2,207-2,761, AIT School Grades Unstandardized n = 77-85, AIT School Grades Standardized n = 76-84. Attrition is excluded because individuals that attrit in their first months are not eligible to take the research-only assessments.

APPENDIX C

VALIDITY SUPPORTING ANALYSES

Table C.1. Bivariate Correlations between the TAPAS Scales and Can-Do Performance-Related Criteria for Tier 1 Soldiers

	Criteria							
	WTBD JKT	MOS-Specific JKT	MOS-Specific PRS	IMT Exam Grade	Graduated IMT without Restart (Academic)	Training Achievement (ALQ)	Training Restarts (ALQ)	Can-Do (PRS)
TAPAS Facets	*n* = 4,425	*n* = 3,725	*n* = 1,332	*n* = 7,238	*n* = 15,739	*n* = 4,546	*n* = 4,556	*n* = 1,585
Achievement	**.06**	**.04**	.04	**.07**	-.01	**.09**	**-.09**	**.07**
Adjustment[a]	**.07**	**.05**	.00	.01	-.01	.00	**-.05**	-.02
Attention Seeking	**.05**	.02	-.02	.00	.01	**.04**	**-.05**	.00
Cooperation	**-.03**	-.03	-.02	.00	.00	-.03	.01	-.01
Dominance	**.05**	.00	.02	.02	.01	**.11**	**-.11**	.05
Even Tempered	**.04**	.02	-.02	**.02**	-.02	**-.05**	.02	.00
Generosity	**-.04**	**-.03**	**-.06**	**-.03**	.00	-.01	**.05**	-.05
Intellectual Efficiency	**.23**	**.18**	-.03	**.13**	-.01	-.02	**-.09**	.00
Non-delinquency	-.02	-.03	-.01	**.04**	-.01	-.01	.02	.01
Optimism	.01	-.02	**.06**	.02	.00	.01	**-.04**	**.06**
Order	**-.07**	**-.07**	-.02	**-.03**	-.01	**.05**	.03	.00
Physical Conditioning	.02	-.03	**.06**	.00	**.06**	**.13**	**-.17**	**.09**
Self Control[a]	-.01	-.01	.01	.01	**-.02**	.01	.01	.01
Sociability	**-.08**	**-.09**	.01	**-.06**	.01	**.05**	-.02	.00
Tolerance	**-.03**	-.03	-.02	**-.03**	**-.02**	.01	**.08**	-.01

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 1,282-15,114.

Table C.2. Bivariate Correlations between the TAPAS Scales and Will-Do Performance-Related Criteria for Tier 1 Soldiers

	Criteria					
TAPAS Facets	Effort and Discipline (PRS) *n* = 1,603	Exhibiting Fitness & Bearing (PRS) *n* = 1,595	Work with Others (PRS) *n* = 1,592	Last APFT Score (ALQ) *n* = 4,504	Disciplinary Incidents (ALQ) *n* = 2,962	Overall Performance (PRS) *n* = 1,591
Achievement	**.10**	**.08**	**.07**	**.09**	**-.06**	**.10**
Adjustment[a]	-.02	-.01	-.04	.00	-.01	-.03
Attention Seeking	.01	.05	.03	**.06**	.01	.01
Cooperation	-.05	-.02	-.03	-.01	.00	-.04
Dominance	**.05**	**.06**	.04	**.13**	**-.04**	.04
Even Tempered	.03	.02	.01	**-.07**	-.02	.01
Generosity	-.04	-.04	**-.05**	.00	-.01	.01
Intellectual Efficiency	.00	-.01	.00	**.05**	-.02	.04
Non-delinquency	.00	-.03	.01	**-.05**	**-.04**	-.01
Optimism	**.05**	.04	**.06**	**.04**	.00	**.06**
Order	.01	.01	.01	.02	.00	.00
Physical Conditioning	**.08**	**.15**	**.08**	**.28**	**-.08**	**.10**
Self Control[a]	.02	-.01	-.02	-.02	**-.04**	.00
Sociability	-.01	.01	.01	**.04**	.03	-.01
Tolerance	-.01	-.01	-.02	.02	.01	.03

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).
[a] Adjustment and Self-Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 1,533-4,387.

Table C.3. Bivariate Correlations between the TAPAS Scales and Retention-Related Criteria for Tier 1 Soldiers

	Criteria								
	Affective Commitment (ALQ)	Attrition Cognitions (ALQ)	Adjustment to Army Life (ALQ)	Commit and Adjust (PRS)	Army Fit (ALQ)	MOS Fit (ALQ)	3-Month Attrition[b]	6-Month Attrition[b]	9-Month Attrition[b]
TAPAS Facets	*n* = 4,556	*n* = 4,556	*n* = 4,556	*n* = 1,600	*n* = 4,556	*n* = 4,556	*n* = 23,526	*n* = 17,964	*n* = 10,667
Achievement	**.14**	**-.14**	**.15**	**.07**	**.15**	**.11**	.00	-.01	-.01
Adjustment[a]	**-.03**	-.02	**.09**	-.02	.00	.03	**-.02**	**-.02**	-.02
Attention Seeking	**.06**	**-.04**	**.05**	.01	**.04**	**.04**	**-.01**	**-.02**	-.02
Cooperation	.01	-.01	-.01	-.03	.00	-.02	.00	.00	-.01
Dominance	**.11**	**-.09**	**.14**	.04	**.12**	**.08**	.00	**-.02**	-.01
Even Tempered	.01	**-.04**	**.04**	.00	.02	-.01	-.01	-.01	**-.02**
Generosity	**.06**	**-.04**	-.01	**-.06**	**.06**	.02	**.03**	**.03**	**.03**
Intellectual Efficiency	-.01	**-.05**	**.12**	-.01	**.03**	**.03**	.00	-.01	-.01
Non-delinquency	**.05**	**-.03**	.02	.00	**.04**	.01	.01	.01	.01
Optimism	**.06**	**-.06**	**.11**	**.06**	**.07**	**.06**	**-.01**	**-.03**	**-.03**
Order	.01	.02	-.01	.00	.01	**-.04**	**.01**	**.02**	**.03**
Physical Conditioning	.03	**-.05**	**.12**	**.08**	**.05**	**.09**	**-.04**	**-.07**	**-.06**
Self Control[a]	**.03**	-.02	.03	-.01	**.03**	-.02	.00	.00	.00
Sociability	**.04**	.00	.02	-.01	.02	**.06**	.00	.00	.00
Tolerance	**.05**	**-.04**	**.03**	-.01	**.05**	.01	.00	.01	.01

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 1, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are smaller, ranging from 1,541-23,101.

[b] Attrition results include Regular Army Soldiers only.

Table C.4. Bivariate Correlations between the TAPAS Scales and Can-Do Performance-Related Criteria for Tier 2 Soldiers

	Criteria							
	WTBD JKT	MOS-Specific JKT	MOS-Specific PRS	IMT Exam Grade	Graduated IMT without Restart (Academic)	Training Achievement (ALQ)	Training Restarts (ALQ)	Can-Do (PRS)
TAPAS Facets	*n* = 109	*n* = 89	*n* = 31	*n* = 163	*n* = 452	*n* = 112	*n* = 112	*n* = 36
Achievement	.17	.15	-.15	.10	.02	.07	**-.20**	-.17
Adjustment[a]	-.04	-.13	-.18	.06	-.03	**-.25**	-.08	**-.37**
Attention Seeking	.10	-.02	-.08	.01	.04	-.15	.01	-.08
Cooperation	.05	.09	-.25	.12	.01	.08	-.12	-.24
Dominance	.10	.04	-.01	-.01	-.03	.12	-.05	.05
Even Tempered	**.22**	.14	-.22	-.08	**-.10**	.05	.00	-.17
Generosity	-.01	.16	**-.44**	.04	-.02	-.10	-.04	-.24
Intellectual Efficiency	.13	.15	-.07	**.28**	.07	-.07	-.04	-.18
Non-delinquency	**.19**	.19	-.11	**.18**	-.01	-.02	-.13	-.06
Optimism	-.08	-.08	.22	.05	-.04	.04	-.16	.12
Order	-.04	.07	-.24	-.06	.02	-.05	.09	-.08
Physical Conditioning	-.05	-.08	**.39**	-.01	-.02	.08	-.17	.23
Self Control[a]	.10	.01	-.08	.06	-.03	.09	-.07	-.14
Sociability	.12	.07	.34	-.03	-.01	-.04	-.02	.33
Tolerance	-.13	-.13	-.22	.02	-.03	-.14	.03	-.08

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 2, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales the same or are smaller, ranging from 31-436.

Table C.5. Bivariate Correlations between the TAPAS Scales and Will-Do Performance-Related Criteria for Tier 2 Soldiers

	Criteria					
	Effort and Discipline (PRS)	Exhibiting Fitness & Bearing (PRS)	Work with Others (PRS)	Last APFT Score (ALQ)	Disciplinary Incidents (ALQ)	Overall Performance (PRS)
TAPAS Facets	*n* = 36	*n* = 36	*n* = 36	*n* = 109	*n* = 67	*n* = 35
Achievement	-.05	-.18	-.17	.14	-.14	-.03
Adjustment[a]	-.30	-.22	**-.39**	.11	-.04	-.18
Attention Seeking	-.09	-.04	.06	.06	-.03	.01
Cooperation	-.08	-.20	-.26	.12	-.11	**-.33**
Dominance	-.10	-.13	.02	.16	-.18	.05
Even Tempered	-.04	.02	-.10	-.14	.17	.04
Generosity	-.24	-.23	-.26	-.02	-.05	-.05
Intellectual Efficiency	-.23	-.18	-.11	-.04	.05	-.15
Non-delinquency	.25	-.05	.16	.11	-.04	-.14
Optimism	.29	.17	.04	**.40**	-.21	-.12
Order	.11	.03	-.04	.03	.24	.07
Physical Conditioning	.14	.18	.17	.14	-.20	**.43**
Self Control[a]	.00	-.04	-.12	.08	-.03	-.12
Sociability	.19	.29	.24	.05	-.11	.13
Tolerance	-.07	-.10	-.03	.05	-.08	-.11

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 2, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are the same or smaller, ranging from 35-108.

Table C.6. Bivariate Correlations between the TAPAS Scales and Retention-Related Criteria for Tier 2 Soldiers

	Criteria								
	Affective Commitment (ALQ)	Attrition Cognitions (ALQ)	Adjustment to Army Life (ALQ)	Commit and Adjust (PRS)	Army Fit (ALQ)	MOS Fit (ALQ)	3-Month Attrition[b]	6-Month Attrition[b]	9-Month Attrition[b]
TAPAS Facets	*n* = 112	*n* = 112	*n* = 112	*n* = 36	*n* = 112	*n* = 112	*n* = 272	*n* = 221	*n* = 157
Achievement	.163	-.090	**.217**	-.159	.173	.043	.044	.038	.026
Adjustment[a]	-.123	.124	.117	**-.420**	-.088	-.093	-.020	.045	.053
Attention Seeking	.094	-.080	.092	-.101	.094	.011	-.116	-.042	-.106
Cooperation	-.037	.085	.016	-.165	-.088	-.051	-.018	-.048	-.126
Dominance	-.016	.099	.174	-.021	.034	.032	.058	-.005	.001
Even Tempered	.033	-.133	.034	-.195	.005	.098	-.011	.007	-.028
Generosity	.042	.003	.055	-.246	.054	.021	-.012	-.001	-.035
Intellectual Efficiency	-.075	.072	-.010	-.201	-.110	-.008	.019	-.034	-.077
Non-delinquency	.008	.030	.117	.112	.026	.077	.013	-.004	-.032
Optimism	.134	-.143	**.202**	.008	.125	-.015	.005	.045	.143
Order	.004	-.013	.121	-.017	.029	-.008	.015	.012	.028
Physical Conditioning	.165	-.138	**.299**	.130	**.194**	.097	.055	.099	.145
Self Control[a]	-.001	-.040	.051	-.185	-.020	-.081	**.192**	**.145**	**.158**
Sociability	.043	-.011	.059	.251	.031	.035	**-.195**	-.119	-.130
Tolerance	-.051	.110	-.050	-.123	-.100	-.047	-.071	-.041	-.045

Note. AFQT = Armed Forces Qualification Test, TAPAS = Tailored Adaptive Personality Assessment System. ALQ = Army Life Questionnaire. JKT = Job Knowledge Test. PRS = Performance Ratings Scales. Results are limited to non-prior service, Education Tier 2, AFQT Category IV and above Soldiers. Estimates in bold were statistically significant, $p < .05$ (two-tailed).

[a] Adjustment and Self Control were included in the TAPAS 15-dimension versions (i.e., static and CAT) only. Sample sizes for these scales are the same or smaller, ranging from 36-270.

[b] Attrition results include Regular Army Soldiers only.